THE UNIVERSITY
of EDINBURGH

Targeting tyrosine: a catch-and-release approach to protein modification

Christopher Allan

THE UNIVERSITY *of* EDINBURGH

A thesis submitted for the degree of

Doctor of Philosophy

The University of Edinburgh

2018

Declaration

This thesis is submitted in part fulfilment of the requirement for the degree of Doctor of Philosophy at the University of Edinburgh. Unless otherwise stated, the work described in this thesis is original and has not been submitted previously in whole, or in part for any degree or qualification at this, or any other university. In accordance with the dissertation regulations as specified by The University of Edinburgh, this thesis does not exceed 100,000 words in length.

Christopher Allan

August 2018

Acknowledgements

I would like to take this opportunity to sincerely thank several people who have contributed to my development over the course of this PhD, not only as a researcher but also in a personal capacity.

I would firstly like to thank my supervisor Prof. Alison N. Hulme, who has taught me a great deal of chemistry, allowed me significant freedom in the work I have attempted, and provided enormous support and motivation with her infectious enthusiasm for research. I would also like to thank my second supervisor Prof. Rory R. Duncan for very helpful discussions about the application of my research in a Biological setting, and support through each annual submission period. I would like to acknowledge the financial support I have received from the Engineering and Physical Sciences Research Council and the Medical Research Council which has allowed me to undertake this work.

My time in the Hulme group has overlapped with many influential characters, who I sincerely thank, but are too numerous to name. Needless to say, I consider myself very fortunate to have coincided with a group of people who could be at once so full of knowledge and so full of nonsense. Being part of this group has kept me entertained and motivated in no small measure. I would like to specifically thank Christina and Mirek who came to work on undergraduate projects with me, and brought considerable skill and enthusiasm to the job.

For assistance in MS and spending a lot of time teaching me and others how to use the MS facilities in the department, I would like to thank Dr C. Logan MacKay and Dr Faye Cruickshank. For assistance in NMR, I would like to thank Dr Lorna Murray and Mr Juraj Bella.

I would like to express my deep-felt thanks and gratitude to my family for giving me all the support and all the freedom that has allowed me to pursue this PhD. Thanks also to my future in-laws who have been enormously supportive, and especially to the newest member, Mallie, who has come along and changed everything.

Finally and most importantly, I would like to thank my fiancé Gillian. Throughout my PhD you have been an amazing support and the perfect distraction; I feel incredibly lucky to have you in my corner. Thank you for all of the proof-reading, for being a practice audience member, and for laughing at all of the chemistry terminology which now creeps into my everyday speech.

Abstract

Protein modification is an essential tool in Chemical Biology, allowing a functional biomolecule to be equipped with a small molecule tag or label. However, as proteins are constructed from a limited palette of around 20 canonical amino acids, achieving selective modification can be problematic. Previously reported methods for protein modification will be discussed in **Chapter 1**; these often rely on alteration of the protein sequence to introduce a uniquely reactive (often non-canonical) amino acid which may then be covalently modified in a bioorthogonal manner. An alternative approach is to identify a uniquely reactive site within the native protein sequence, such as the protein N-terminus or the reactive side chain of an amino acid with low frequency, and modify this using selective chemistry.

In this project, modification of a native sequence protein was achieved by targeting a low abundance residue, tyrosine (Tyr), in a selective manner. Tyr was identified as the ideal candidate as it displays only ~3% frequency in the proteome and, due to its electron-rich aryl ring, it can be selectively modified by electrophilic aromatic substitution. Using a diazonium salt as the tuned electrophile, modification results in formation of an azobenzene motif which may be orthogonally cleaved under mild reducing conditions. The resulting cleavage product bears an *o*-aminophenol modification on the Tyr side chain, which can then be conjugated to a fluorescent label using established chemistry. This system has been developed on a solid-phase platform to give further control over the extent of modification achieved.

In **Chapter 2**, the component parts of this method are developed through reactions performed in-solution on small molecule substrates. In **Chapter 3**, this work is then moved onto a solid-phase resin in order to 'catch-and-release' small molecule and peptide substrates. Finally in **Chapter 4**, the resin-based catch-and-release system is optimised for use in protein modification, and analysis of the modification site is explored.

Lay Abstract

Proteins are complex functional biomolecules which serve a range of diverse roles in Nature. From the perspective of a researcher, the recruitment of these biomolecules to help answer research questions is highly desirable, and generally this requires labelling of the protein using Chemical or Biological techniques. While proteins are highly diverse in composition and structure, they are all constructed from the same pool of around 20 amino acids. Some of these amino acids bear a side chain which has a reactive group, allowing modification of this building block with a synthetic label such as a fluorescent group.

The work contained in this thesis describes a new method for protein labelling which targets the natural amino acid tyrosine. Tyrosine is an appropriate target as it bears a side chain which shows nuanced reactivity, while its relatively low abundance and moderate surface-exposure offer the potential for selective labelling when more than one tyrosine is present in the protein. Using a solid-phase resin to 'catch' the protein *via* a reactive tyrosine residue, and release it in a modified form, a protein substrate has been successfully labelled with a fluorescent group, and this type of protein bioconjugate could be used in future as a probe for fluorescence microscopy.

Table of Contents

Chapter 1 Introduction: Chemical Biology approaches to protein modification

1.1 Overview

Biomolecules and natural products have proven a rich source of inspiration for various sub-disciplines in Chemistry and Biology, including Medicinal Chemistry, Materials Chemistry and Chemical Biology. Through millions of years of evolution, Nature has developed functional molecules which display exquisite binding specificity and serve roles including catalysis, structural organisation and signalling. These functions are of great interest to the modern day chemist, and through careful modification of native substrates, the researcher can endow these functional biomolecules with an orthogonal functionality which suits their needs.

The work presented in this thesis is particularly concerned with bioconjugation, where a biomolecule, such as a peptide, protein or nucleic acid, is labelled with a functional group. These labels may be synthetic small molecules such as a ligand, drug or reporter group, or biological labels such as carbohydrates, nucleic acids or lipids. In addition, macromolecular labels may be employed, such as functional proteins or polymers; or bioconjugation chemistry may be used for attachment of a biomolecule to an activated surface.[1] Depending on the application, the choice of biomolecular substrate and functional label will differ, and common examples of these will be explored. For the current project fluorescent reporters are the labels of interest, and through bioconjugation with a protein substrate, a protein-based fluorescent probe is formed. For this reason, nucleic acids as biomolecular substrates or labels will not be discussed.

While various methods have been reported to achieve covalent attachment to a biomolecule, achieving chemoselectivity and site-specificity remains challenging. This has given rise to the development of a toolbox of protein modification strategies, each of which are best suited to specific applications. The most common examples of these toolbox techniques will be discussed, incorporating both ‘Chemistry’ and ‘Biology’ approaches.

1.2 Biomolecular substrates

The specific choice of biomolecular substrate depends on the application at hand, but these biomolecules generally act as the targeting group for the finished bioconjugate. Due to their specific binding capabilities, antibodies and their derivatives are often used. Full-length antibodies are used for the production of targeted therapies such as antibody–drug conjugates (ADCs), as well as for targeted fluorescent dyes used in microscopy. They exhibit very strong and very specific binding to an antigen, and are composed of two heavy chains and two light

chains organised in a distinctive 'Y' shape with the variable binding region at the tip of each arm (Figure 1-1). The majority of the antibody structure consists of highly conserved protein sequences (constant regions: C_H1, C_H2, C_H3, C_L), and antibody labelling is commonly achieved through modification of these regions (often this involves targeting conserved disulfide bridges).

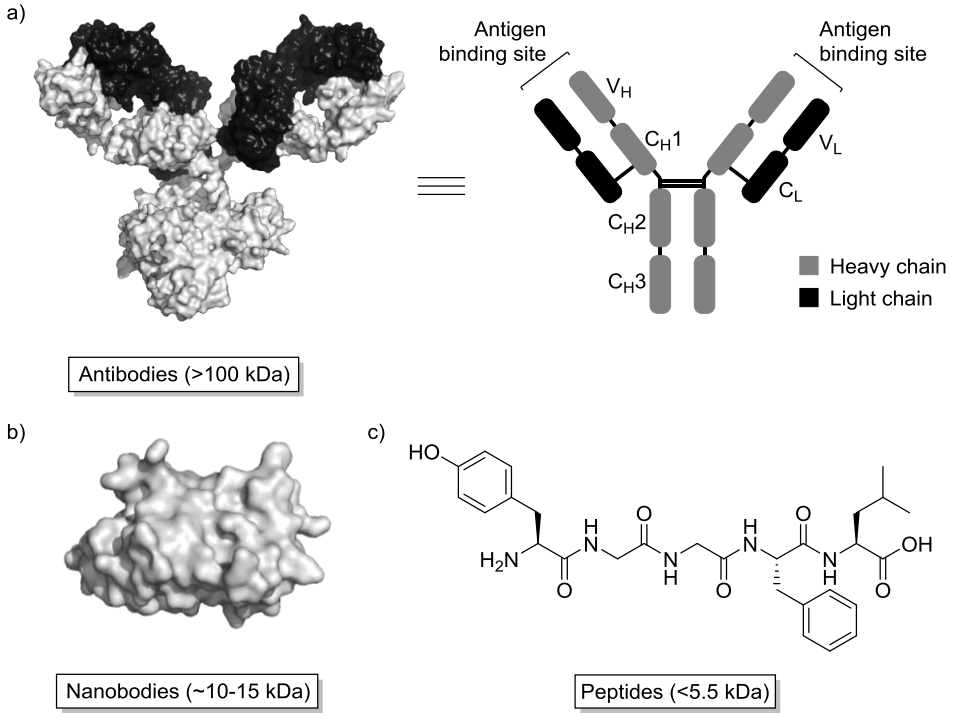


Figure 1-1 *Biomolecular substrates used for bioconjugation. (a) Antibodies exhibit very strong and selective binding, and have therefore been used in the development of ADCs and fluorescent probes (V: variable domain; C: constant domain). (b) Nanobodies are the isolated V_H domain from a heavy chain antibody, and retain strong binding affinity but at roughly one tenth of the size of an antibody. (c) Peptides are small in size and easier to synthesise and modify than protein substrates, but a labelled peptide may lack the specificity of a protein-based probe. (PDB ID's: 1IGY, 3OGO).*

While antibodies are employed in a wide range of applications, they are large proteins at >100 kDa, which make them difficult to handle and modify. In fact, as the binding interaction occurs at a distal region of the protein, the majority of the antibody bulk is not involved in binding. A more compact and atom-efficient alternative to antibodies is found in their nanobody counterpart. These antibody derivatives may be obtained from camelid or shark species, which naturally produce antibodies containing only a heavy chain. The variable binding region, V_H, may be isolated by enzymatic digestion of the intact antibody.[2] The nanobody is around one tenth of the size of an antibody at 10–15 kDa, yet retains a strong antigen-specific binding affinity, and is therefore an appealing biomolecular substrate for the production of protein based-probes.

Among the benefits of using peptides as biomolecular substrates is the fact that many potent peptide species are small enough to be synthetically achievable, allowing the production of relatively large quantities without relying on protein expression and purification. In addition, smaller peptides are likely to tolerate harsh labelling conditions, as they do not exhibit the higher order structure of a protein. If a substrate is synthesised using conventional solid-phase peptide synthesis, this would also provide excellent regulation over labelling, as the sequence is assembled one amino acid at a time, providing temporal control over site-specificity, while there is also orthogonal protection of potentially competitive reactive groups. The biomolecular labelling methods reviewed in this Introduction will focus on protein-compatible strategies, however a recent review by Baran *et al.* describes a wider range of methods that are applicable to peptide substrates.[3]

1.3 Functional labels

In conjunction with the choice of biomolecular substrate, the appropriate functional label is selected based on the application of the final bioconjugate. Many functional labels are described in the literature, which may equip a biomolecular substrate with a range of functionalities (Figure 1-2).

In Medicinal Chemistry, the bioconjugation of antibodies with cytotoxic drug compounds *via* either a cleavable or non-cleavable linker has been used in the production of ADCs. This allows targeted delivery and release of the cytotoxic payload to the desired site, and to date, four ADCs have received approval from the Federal Drug Administration (FDA) for clinical use, while many more are currently in development.[4] Another application of bioconjugation in Medicinal Chemistry involves the attachment of hydrophilic polymers [most often composed of polyethylene glycol (PEG)] to a therapeutic peptide or protein, which has been shown in certain cases to increase blood circulation time, increase solubility and decrease immunogenicity.[5]

In Materials Chemistry, covalent attachment of a protein to a solid support has been used for biophysical assays including Surface Plasmon Resonance (SPR), and in the development of separation materials such as (strept)avidin resins, which can be used for the specific enrichment of biotin-labelled molecules of interest.[6]

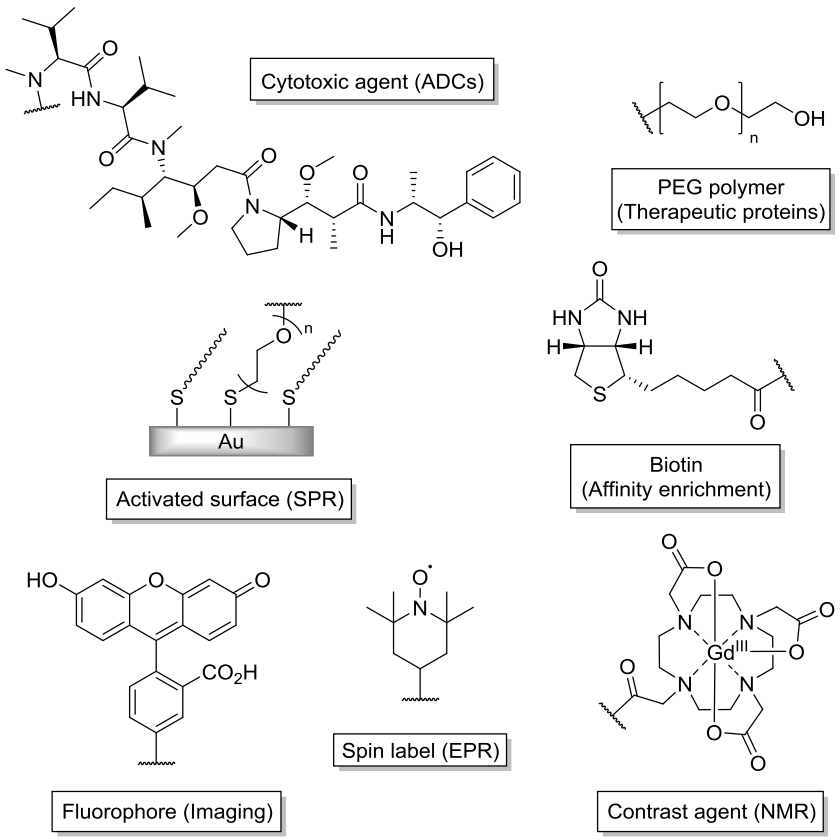


Figure 1-2 *Selected functional labels used for bioconjugation. Depending on the application, the biomolecular substrate may be equipped with a variety of functional labels, ranging from cytotoxic compounds to reporter molecules to affinity labels such as biotin. Alternatively, protein labelling chemistry may be used to immobilise the biomolecule on a solid-phase for use in a biophysical assay such as surface plasmon resonance (SPR).*

Finally, in Chemical Biology, protein modification has allowed the study of native systems using proteins and peptides which have been equipped with reporter molecules, including fluorescent groups, spin labels and NMR contrast agents. Alternatively, attachment of a biotin label allows enrichment of a protein sample, or detection with fluorophore-labelled (strept)avidin.

The methods to produce these bioconjugates are numerous and varied, but can generally be classed as methods which alter the protein sequence, or methods which selectively target a site in the native sequence protein.

1.4 Methods for protein modification

1.4.1 Alteration of the protein sequence

A genetically engineered or 'Biological' approach to protein modification involves the alteration of a protein's native sequence to introduce a fused functional protein, a peptide recognition sequence or an orthogonally reactive natural or non-natural amino acid. These methods are useful for *in vivo* protein labelling, as the modified species are expressed in cells. However, the design and execution of these techniques require bespoke biochemical methods and optimisation for each protein substrate of interest, and their use is therefore limited to laboratories with specialist biochemical equipment.

Fusion proteins

The labelling of a biomolecular substrate with a functional label *via* the production of a chimeric fusion protein represents a useful one-step approach to protein labelling. Following the successful cloning and expression of green fluorescent protein (GFP), production of fusion proteins in which the GFP sequence was appended to the native sequence of a protein substrate became an attractive route to the development of protein-based fluorescent probes (Figure 1-3).

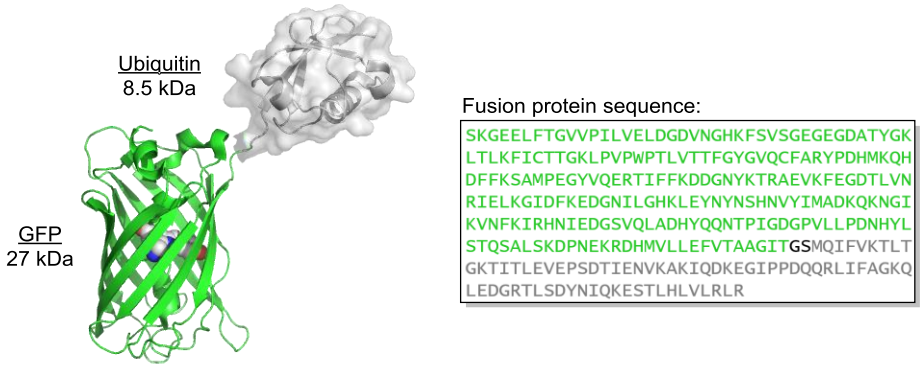


Figure 1-3 *The Ubiquitin-GFP fusion protein. The GFP label (green) is appended to the N-terminus of the ubiquitin native sequence (grey). The two amino acid linker region is shown in the protein sequence in black, the GFP chromophore is shown as spheres. (PDB ID: 3AI5).*[7]

In recent years, the range of genetically encoded fluorescent proteins has expanded to allow multi-colour labelling, which is useful for optical imaging experiments,[8] however the performance of fluorescent proteins in terms of quantum yield and photostability are generally poorer than their synthetic fluorophore counterparts.[9, 10] The benefit of the fusion protein approach is strict control over the stoichiometry and site-specificity of labelling, however the range of labels is limited to those which can be expressed using the protein production machinery. As well as the limited functionality of expressed protein labels, the large size of

these groups compared with small molecule reporters is more likely to perturb the native function of the biomolecular substrate.[11]

In order to expand the range of tolerated labels in the fusion protein approach, self-labelling enzyme technologies have been developed, which functionalise the biomolecular substrate in a two-step procedure. The required fusion protein is first generated, with a self-labelling enzyme encoded at the N- or C-terminal end of the substrate, before selective labelling using the specific substrate recognised by the fused enzyme. The most common examples of this approach involve the use of Halo-tag, SNAP-tag or CLIP-tag (Figure 1-4).[12]

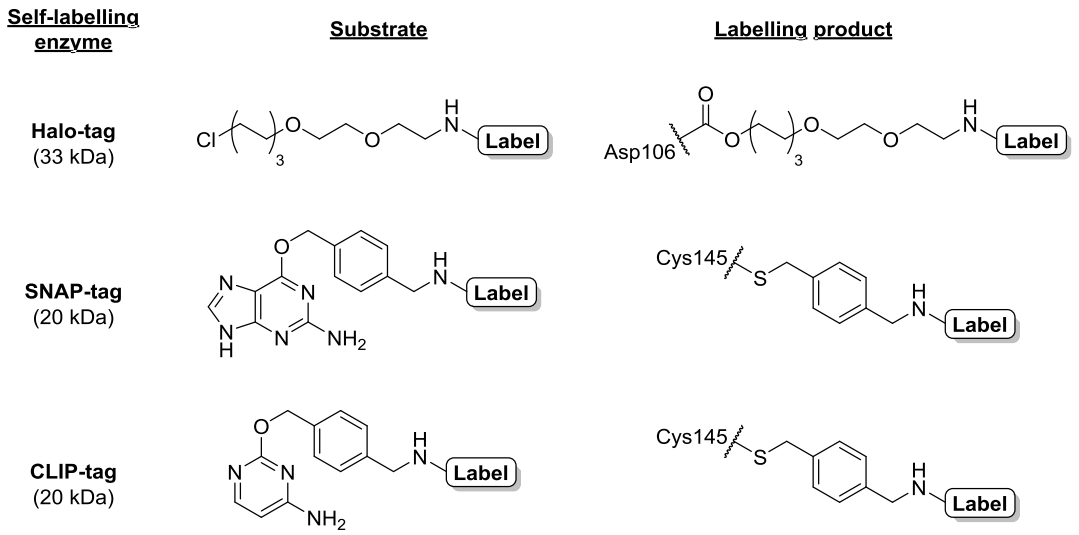


Figure 1-4 *Selected self-labelling enzymes used for the production of self-labelling fusion proteins. Halo-tag is modified by reaction with a halo-alkane substrate, to give the labelled product via ester formation with Asp106 in the Halo-tag sequence.[13] SNAP-tag and CLIP-tag are derived from the same enzyme, but accept different substrates. They are both modified at the active site Cys145 to give a thioether linkage to the label.[12, 14]*

Using self-labelling enzyme incorporation, protein substrates have successfully been labelled with synthetic fluorophores and biotin labels, and this technology has also been used for the immobilisation of a biomolecular substrate onto a solid support.[13, 15, 16] These methods have therefore expanded the range of labels tolerated by a fusion protein approach, while maintaining the site-specificity and stoichiometry. However, the addition of a large (20–40 kDa) biomolecule to the substrate sequence is still required, which may be detrimental to the function and stability of the substrate. Furthermore, genetic manipulation of each target protein is required, which potentially limits the application of these techniques.

Peptide handles

In order to avoid the difficulties associated with the incorporation of a large protein label, peptide handles have been developed which can facilitate protein labelling in a two-step strategy, through enzymatic recognition of a short amino acid sequence. The first example of this approach was reported by Cronan, who found that incorporation of a 75 amino acid sequence into a target protein resulted in the substrate becoming biotinylated *in vivo*, which in turn allowed its purification from cell lysates.[17] Beckett *et al.* later refined this approach by reducing the size of the required recognition peptide to a 14–15 amino acid sequence, which was selectively biotinylated by a biotin ligase enzyme (BirA) (Figure 1-5).[18]

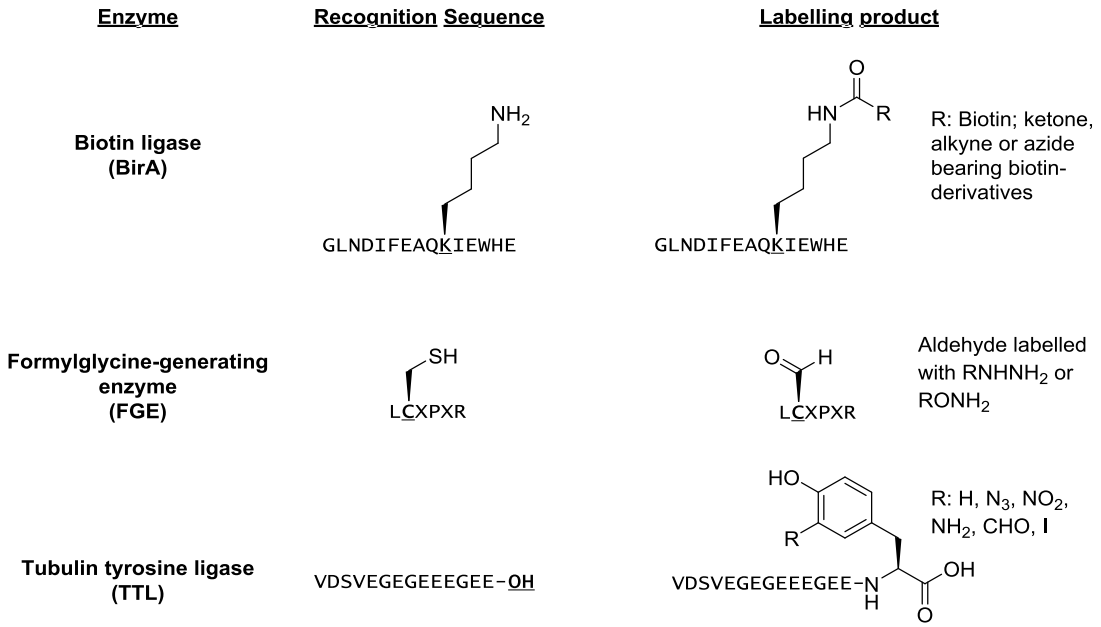


Figure 1-5 *Selected peptide handles used for enzymatic protein labelling strategies. The tolerated substrates of BirA and TTL enzymes were expanded to include azide and alkyne groups, which allowed later functionalisation using click chemistries. FGE results in formation of a reactive aldehyde, which may be functionalised through oxime- or hydrazone-forming reactions. The peptide recognition sequences for BirA and FGE may be incorporated at either the N- or C- termini, however the TTL-specific sequence requires incorporation at the C-terminal.*[19]

As well as its use in affinity purification, this technology has since been used to achieve the site-specific immobilisation of an antibody fragment onto a solid support.[20] Biotin is of course a useful label to attach to a biomolecule, however the BirA labelling technology was further developed by an expanded scope of carboxylic acid substrates which the enzyme could tolerate and transfer onto the peptide Lys residue, including azide and alkyne-bearing derivatives of biotin.[21] Labelling of a peptide handle with a click-functionalised biotin analogue then allows a third step of selective click conjugation which could incorporate a range of functional labels.

Aside from the biotinylating enzyme BirA, a range of enzymes have been used for similar peptide recognition-labelling strategies.

The formylglycine-generating enzyme (FGE) has been employed to recognise a 6 amino acid receptor peptide, and convert a Cys side chain within it to a reactive aldehyde. This approach allows modification of the target protein with a synthetic hydrazine or hydroxylamine-bearing reagent to give a hydrazone or oxime functionalised bioconjugate.

The final peptide handle highlighted is the 14 amino acid sequence which is recognised and modified by tubulin tyrosine ligase (TTL). This peptide sequence (Tub-tag) is labelled by addition of a Tyr-derived residue at the C-terminus, and work by Hackenberger *et al.* highlighted a rather broad substrate tolerance in this system.[22] Incorporation of a reactive azide, iodide and aldehyde *via* the TTL-dependant system provided a route to later covalent labelling with synthetic groups, and ultimately, this system was used to label a functional nanobody with fluorophores, a PEG group and a biotin handle.[23]

Alongside the highlighted peptide handles, commonly used labels include LAP-tag (22 or 13 amino acid sequence, labelled by lipoic acid ligase), LPXTG-tag (5 amino acid sequence, labelled by Sortase A) and ybbr- or S6-tag (11 or 12 amino acid sequence, labelled by phosphopentathienyl transferase).[19]

Peptide tags which do not require enzymatic recognition such as the His-tag (H_6 or H_{10}) or tetracysteine tag (CCXXCC), are also commonly used in biochemistry, as they imbue the target protein with selective reactivity. Specifically, His-tagged proteins are routinely purified by affinity chromatography, as they bind reversibly to Ni^{2+} ions. Meanwhile, the tetracysteine tag has been used to attach turn-on fluorescent probes, due to their specific affinity for a range of biarsenical reagents.[24]

Taking the idea of a small peptide tag to its smallest iteration, incorporation of a single reactive residue to the N- or C- terminus of a target protein sequence, or point-mutation to incorporate a reactive residue elsewhere in the protein chain, has also been reported, e.g. for the introduction of a Cys residue.[25] These small sequence modifications can offer selective modification without the significant burden of a full protein label, and thus are expected to show less impact on the function and stability of the target protein. The incorporation of non-canonical amino acids into the target protein has been shown to offer bioorthogonal reactivity, and can be achieved in several ways.

Non-canonical amino acid incorporation

The incorporation of an unnatural amino acid (UAA) into the sequence of a target protein allows two-step labelling of the target biomolecule, through expression of the modified substrate and then labelling *via* a bioorthogonal handle on the UAA. The incorporation of a range of amino acid derivatives can be achieved in either a residue-specific or site-specific manner, using auxotrophic cell lines or genetic code expansion technologies.[26] In the first example, there is no requirement for genetic modification of the target protein, however the expression system used will be depleted in one or more amino acid substrates. Incorporation of a non-canonical amino acid in place of the native equivalent then relies upon its acceptance by the appropriate tRNA/tRNA synthetase pair, and results in incorporation of the UAA in place of the depleted residue.[27] This technology has been used to incorporate a range of UAAs bearing click handles in place of native methionine residues (Figure 1-6), and the UAAs highlighted were found to be accepted by the native Met-tRNA synthetase and incorporated into the native protein sequence in place of the Met residue, albeit at a slower rate than the native substrate.[27]

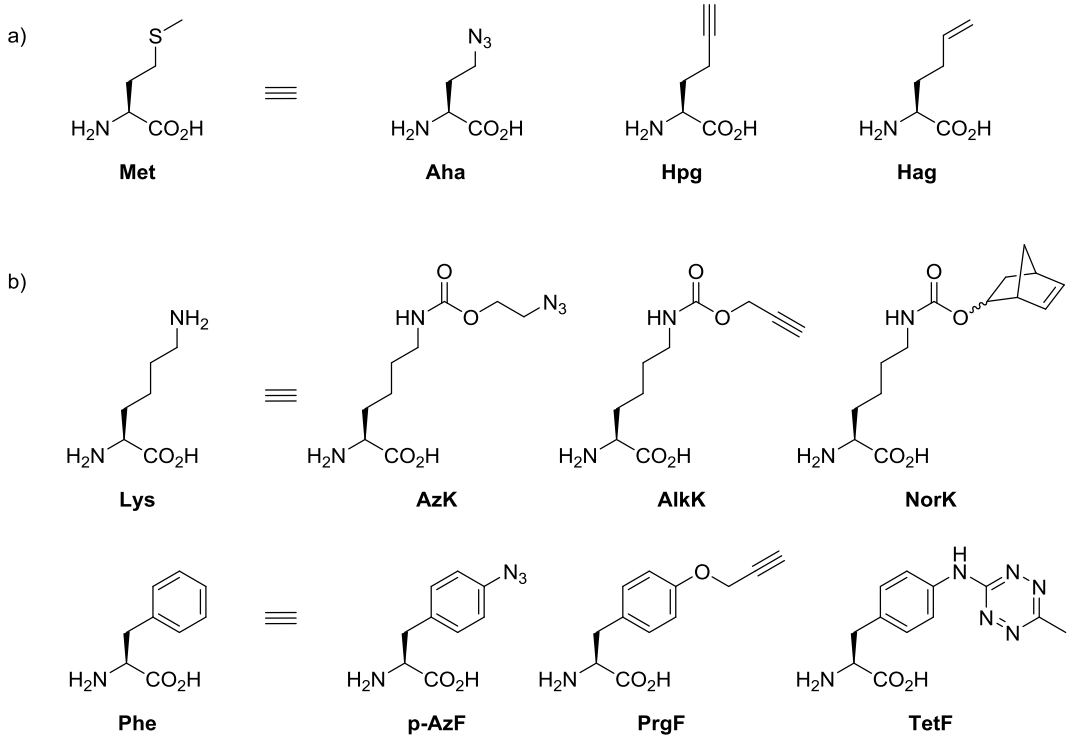


Figure 1-6 *Selected non-canonical amino acids used in the auxotrophic cell line and genetic code expansion approaches.(a) Met-analogues bearing click handles are shown, these are all tolerated by the native Met-tRNA synthetase in the auxotrophic cell line approach. (b) Selected UAAs bearing click handles are shown. Lysine (Lys) and phenylalanine (Phe) are common native scaffolds to use for the design of UAAs, which are used in the genetic code expansion approach.*[27]

In the genetic code expansion approach, a blank codon such as the amber stop codon is reassigned as a signal to incorporate the UAA, and this involves the recruitment of a modified tRNA/tRNA synthetase pair which are specific for the synthetic residue.[28] In this way, the site-specific incorporation of a UAA is achieved, and its position in the protein sequence can be modified by introducing the formerly blank codon at different positions on the gene of interest.[28]

Both the auxotrophic cell line and genetic code expansion methods have proved extremely effective at producing functionalised proteins which bear bioorthogonal handles with a high degree of selectivity. However, these approaches require genetic manipulation of the target protein or its assembly components, and can therefore require time-consuming optimisation for each protein substrate of interest. An alternative strategy is to address the issue from a purely 'Chemical' perspective, and develop selective modification strategies which target reactive sites in the native protein target, thus avoiding any alteration of the expressed protein sequence.

1.4.2 Targeting a native site

In a native polypeptide, there is no shortage of potentially reactive sites which could be covalently modified in a protein labelling experiment. In addition to the α-NH_2 of the N-terminus and the terminal CO_2H of the C-terminus, many amino acid side chains bear reactive groups which offer potential as endogenous targets. In contemplating these options, several variables should be considered which may impact the ability to achieve controlled and selective modification, which is preferred for the majority of bioconjugate applications.

The primary consideration when selecting an endogenous target is its reactivity under conditions which will be tolerated by the biomolecular substrate as a whole without causing denaturation. While the compatible conditions of a given substrate will vary depending on its intrinsic stability and solubility, this generally equates to near-neutral pH, a temperature at or below 37 °C, and a low substrate concentration in the micromolar range. Reactive amino acid side chains which are commonly targeted in protein modification experiments are highlighted in Figure 1-7, while the pKa of labile protons are given to indicate the conditions required for modification.

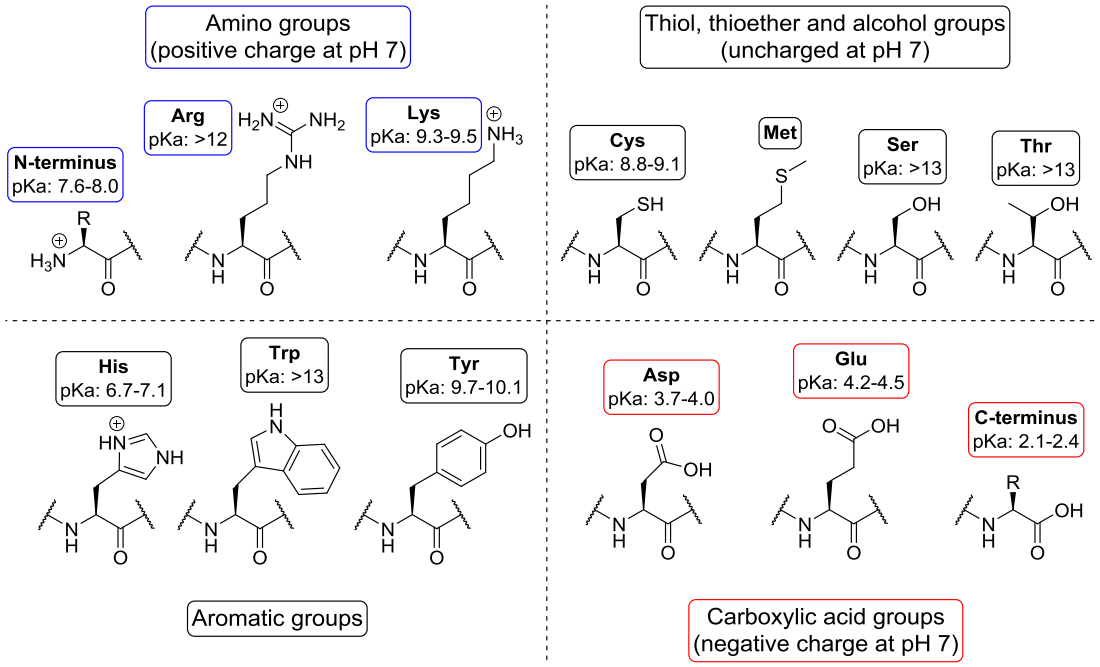


Figure 1-7 *The structures of commonly targeted amino acids are shown, organised by the functional groups present. The pKa's of labile protons are given, as this dictates the reaction conditions required for side chain modification (the pKa of the His imidazolium group lies close to pH 7 and therefore requires slightly basic conditions to favour the imidazole form).*[1]

Through consideration of the pKa of the targeted group, the conditions required for efficient modification are selected. For example, efficient modification of a Cys residue with an electrophile is likely to occur in basic conditions at ~pH 9, where the thiol side chain is largely deprotonated to the more nucleophilic thiolate anion. In cases where there is the potential for cross-reactivity, such as competition between the nucleophilic N-terminal α-NH_2 and the Lys ε-NH_2, careful control of pH may provide selectivity for a single site. In this instance, at ~pH 8.5, the N-terminus is mostly uncharged and nucleophilic, while the Lys amine is more likely to be protonated and therefore not nearly as good a nucleophile. This being said, the local environment of a given residue in a protein sequence will exert some influence over its pKa and therefore reactivity, hence the pKa values in Figure 1-7 are given as a range.

The frequency with which the targeted residue appears in a protein sequence should also be considered when selecting a native target, and the frequency of occurrence of each canonical amino acid in the UniProt/EMBL database is given in Figure 1-8. It is commonplace in protein labelling experiments to use a large excess of reagent due to the high dilution of the substrate and relatively low accessibility of many target residues. For this reason, a strategy which aims to label a highly abundant residue is more likely to result in multiple, non-specific labelling events, giving a stochastic mixture of labelled substrates.

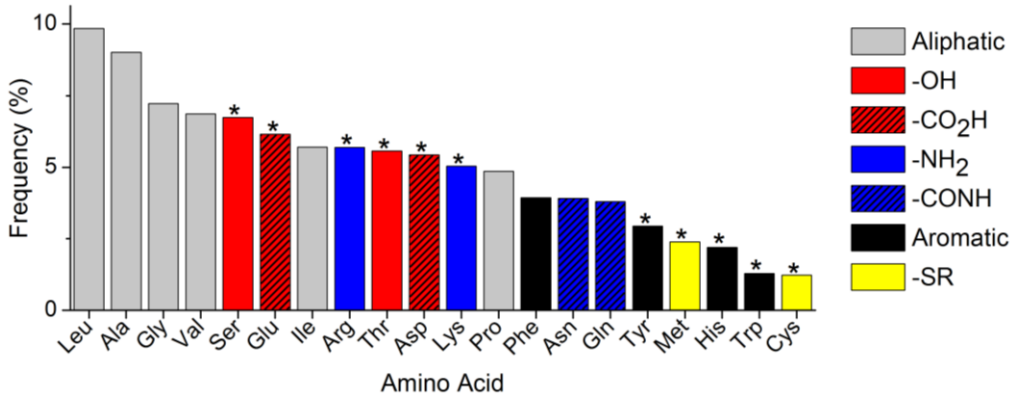


Figure 1-8 *The frequency of each canonical amino acid in the UniProt/EMBL database is shown. The relevant functional group on each amino acid side chain is given in the key, while commonly targeted residues are highlighted with an asterisk (*). (>80 million sequence entries, accessed 26/9/2017).*[29]

For this reason, selective protein modification strategies often target Cys residues, which show the lowest abundance in the gathered data. The scarcity of reactive Cys residues is compounded by their propensity to form intramolecular disulphide bridges, which play a crucial role in preservation of the higher order structure of a protein. Cleavage of disulphides to the corresponding free thiols is possible under reducing conditions, and these may then be re-bridged using a bifunctional linker as a strategy to incorporate a functional label. However, disturbance of the native disulphide bridges may destabilise the protein conformation, while, due to their hydrophobic nature, Cys residues are not generally readily surface accessible, which makes them difficult to address.

Surface-accessibility of a target residue is therefore another variable which should be considered carefully, as it relates to the effective abundance of the targeted group. While there are various examples in Medicinal Chemistry of small molecule ligands binding in deep, cavernous sites within a targeted protein, effective covalent labelling of a protein substrate generally relies upon high surface-exposure of the reactive group, while deeply buried residues are considered 'off limits'. It is well understood that the amino acid composition of the protein interior differs from the surface, with more hydrophobic groups found in the interior where they are shielded from solvent. The reverse is also true, as amino acids with hydrophilic groups on their side chains comprise a larger proportion of the protein surface, where they aid solubility and reduce protein aggregation. This effect is amplified as the size of the protein increases, as the ratio of interior volume to surface area increases in correlation.[30] This is perhaps the reason that the four most abundant amino acids in Figure 1-8 bear hydrophobic side chains.

The hydrophobicity of an amino acid may be assessed in several ways, and various scales of amino acid hydrophobicity exist in the literature. A normalised consensus hydrophobicity scale was published by Eisenberg *et al.*,[31] which combined the findings from empirical and computational assessments of amino acid hydrophobicity, and this is used to rank the 20 canonical amino acids in Figure 1-9 from most hydrophobic to most hydrophilic. This was combined with data published by Chothia *et al.*, who determined the amino acid composition of 37 proteins, and from the related crystal structures, separately calculated the amino acid composition of each protein interior and protein surface.[32] By normalising the protein interior and surface data to the amino acid composition of the whole data-set, the up- or down-regulation of each residue at each site could be established and this has been plotted against each amino acid residue ordered by increasing hydrophilicity in Figure 1-9.

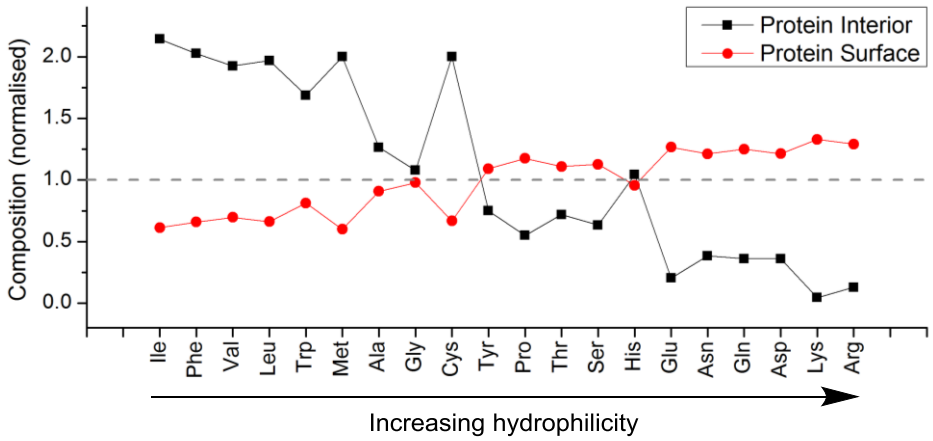


Figure 1-9 *Using data gathered from 37 proteins,[32] the amino acid composition of the protein interior and protein surface was normalised against the total amino acid composition of the dataset. The resulting values are plotted against their respective residues, ordered from most hydrophobic to most hydrophilic following the Eisenberg hydrophobicity scale.[31] These data support the assertion that there is a higher proportion of hydrophobic residues in the protein interior than at the protein surface, and a higher proportion of hydrophilic residues on the protein surface than in the protein interior.*

From the general trend in Figure 1-9, it can be seen that as the hydrophilicity of an amino acid increases, the amino acid represents a higher than average proportion of the protein surface and *vice versa*. It is therefore concluded that amino acids with a hydrophilic side chain will generally exhibit high surface accessibility, and therefore should be easy to address with labelling reagents.

To realise a chemoselective and site-specific labelling method, the reactivity and discrimination potential of the target residue must be balanced. Reactivity relies on the pKa of the reactive group (bearing in mind that this may be affected by the local environment), and

on the surface exposure of the residue (to allow the reactive site to be addressed with an exogenous reagent). To limit the number of labelling events, the reaction conditions should favour only reaction with the target residue, and the overall abundance of this residue and effective abundance based on surface accessibility should limit the number of viable reactive sites.

Protein labelling through covalent modification of an endogenous target has been achieved by targeting many of the amino acids already mentioned, and this research area has been exhaustively reviewed by Koniev and Wagner,[33] Agrawal and Hackenberger,[26] and Algar.[1] The most common targets such as Cys, Lys, the N- and C- termini and Tyr are discussed in subsequent sections, and several recent advances will be highlighted.

Cysteine

Cysteine is an appealing native target due to its high nucleophilicity, low abundance and moderate surface exposure. Thanks to the pKa of the thiol side chain (pKa ~9), Cys is a highly reactive nucleophile in basic conditions, where it can form a thiolate anion. However, at neutral pH, the protonated thiol group is also reactive, whereas many competitive nucleophiles are not (e.g. Lys bears a positive charge at pH 7). The tendency of Cys residues to form structurally important disulphide bridges further lowers the abundance of reactive targets, to the extent that a suitable free Cys is not always present in a native protein structure. In order to address this, point mutation to introduce a reactive Cys, or reduction of an existing disulphide bridge may be a prior requirement to application of the Cys-labelling chemistries which will now be discussed.

In contrast to the hard nucleophilic character of the thiolate anion, the thiol group is a soft nucleophile capable of Michael addition to an electron-poor alkene. The most prominent example of this kind of Cys modification is in the use of maleimide-functionalised labels (Figure 1-10). The maleimide group allows covalent functionalisation of a Cys residue at near-neutral pH to give, in the first instance, a thiosuccinimide-labeled residue. However, the initial Michael addition of the sulfhydryl group to the maleimide is reversible *via* a retro-Michael mechanism, resulting in thiol exchange or elimination of the unmodified Cys.[34] By addition of a basic group adjacent to the maleimide in the Cys-targeting label, or by performing the reaction in basic conditions, hydrolytic ring-opening of the thiosuccinimide is favoured, which forms a more stable covalent linkage.[35, 36] Maleimide Cys labelling has played an important role in the development of ADCs, where conserved Cys residues in the antibody brentuximab

were labelled with the maleimide-functionalised antimitotic peptide derivative monomethyl auristatin E, to give brentuximab vedotin, one of the first ADCs to receive FDA approval.[34]

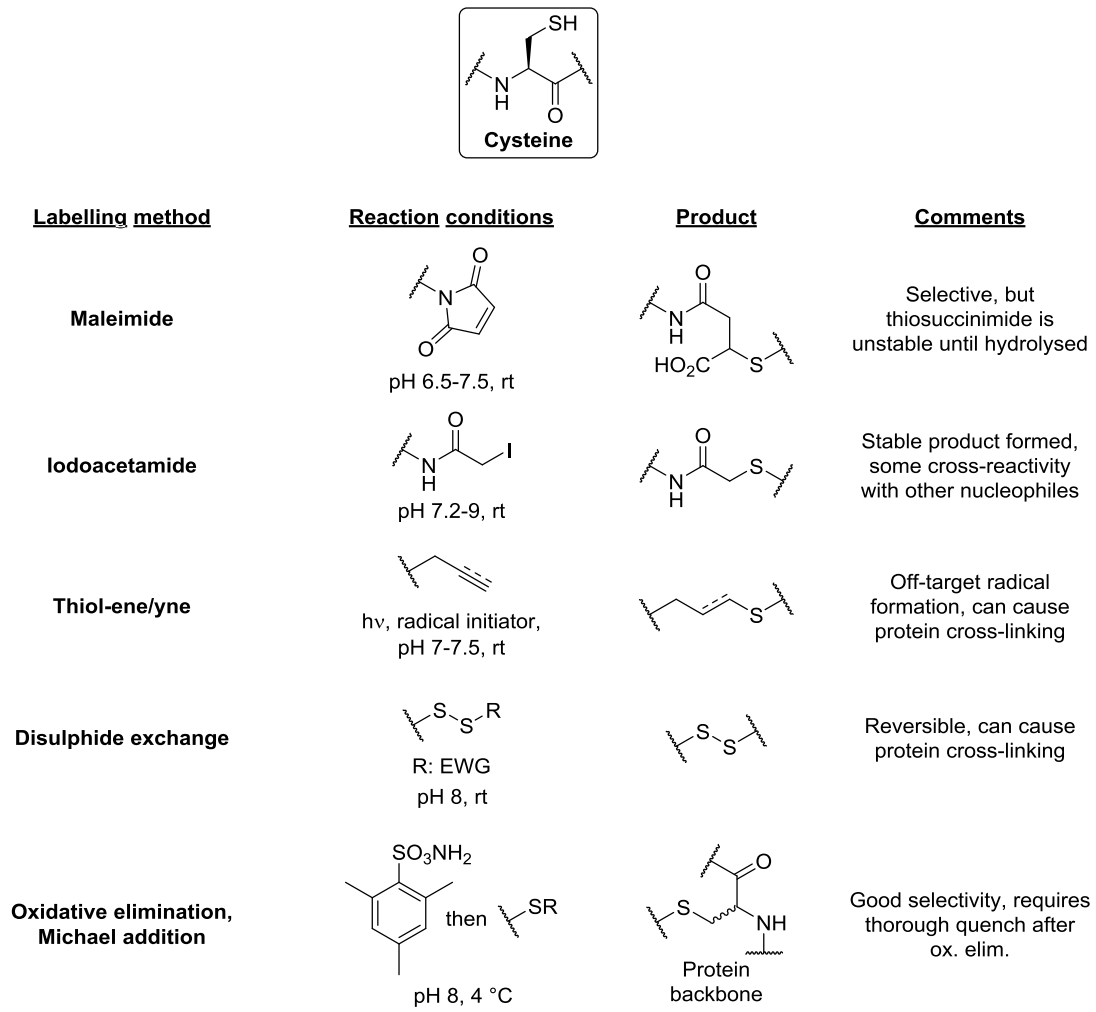


Figure 1-10 *Selected methods for the modification of a Cys residue in a protein substrate.*

Contemporary use of iodoacetamide is commonly confined to the capping of Cys residues which have been generated by disulphide reduction for MS studies.[37] However, historically the α-halocarbonyls have provided an alternative Cys-targeting strategy to maleimides. While these reagents are generally less reactive to the Cys thiol group, this has allowed the reported selective labelling of Cys residues which have a lower pKa due to their existence in a basic local environment.[38] The use of α-halocarbonyls has largely been superseded by the advent of maleimide chemistry and the variations of this group that may be used.

Thiol-ene and thiol-yne reactions are interesting examples for protein labelling as they employ radical mechanisms which require initiation with a radical-generating catalyst and irradiation with UV light. While thiols and alkenes/alkynes are well tuned to react with each other, the

generation of free radicals in a biological sample is likely to cause off-target modifications. Even within the desired labelling reaction, the initial product is a radical species which requires quenching or else it is likely to propagate and react with another partner. In the thiol-yne example, the quenched product contains an alkene, which may be further modified by thiol-ene chemistry. In some respects radical-driven processes are well suited to protein labelling applications, which are generally performed at high dilution thus limiting the radical cross-linking of protein substrates. However, there are photoactive proteins such as fluorescent proteins, which may be adversely affected by irradiation with low wavelength light.[33]

Disulphide exchange is another common Cys modifying approach, and involves the use of a disulphide labelling reagent which contains a motif which will make it a good leaving group. A good example is the use of Ellman's reagent, which is a disulphide species bearing electron-poor aromatic groups on either side of the disulphide bond.[39] These groups facilitate thiol exchange with the Cys thiol group to form a new disulphide, which again bears a good leaving group on one side. This activates the Cys residue for further disulfide exchange with another thiol to give a disulphide labelled substrate.[1] The obvious drawback of this approach is the lability of the linker which is prone to cleavage by reducing agents, however this may be preferred for certain applications, where the reversible labelling and restoration of the native protein can be achieved.[33, 39]

The final approach which has been highlighted involves the conversion of a native Cys residue to dehydroalanine (Dha), which allows subsequent labelling with a thiol reagent *via* Michael addition. Several reagents have been reported which convert Cys to Dha, however the highlighted example is the use of *o*-mesitylenesulfonylhydroxylamine (MSH), which works by successive amination of the Cys thiol group, before oxidative elimination to give the Dha product and a $S(NH_2)_2$ by-product.[40] Davies *et al.* have developed this chemistry, and used the Michael addition of thiols to the generated Dha to mimic phosphorylation and glycosylation PTMs which occur in nature.[41] MSH is known to exhibit some cross reactivity with Met residues, however this is reversible and so reaction at the Cys side chain is preferred. It should be noted that in conversion of the Cys side chain to an alkene and subsequent Michael addition, the stereochemistry at the α-carbon is scrambled.[26]

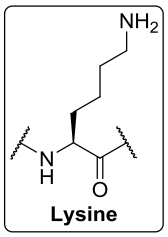
Lysine

As well as showing moderate abundance, Lys bears a hydrophilic side chain, as the ε-NH_2 is predominantly protonated at physiological pH. In fact, Lys is the second most hydrophilic amino acid, behind Arg, and therefore the majority of Lys residues in a protein sequence are

likely to be found on the surface. With this in mind, it should be considered that some Lys modification strategies result in a loss of charge (e.g. by conversion of the Lys amine to an amide) which may be detrimental to the solubility of the protein. Lys is a commonly targeted residue for protein modification, although often in a stochastic labelling approach due to the abundance and accessibility of the reactive group. This being said, a number of recent reports have addressed Lys labelling in a selective manner, and these will be discussed alongside some less selective approaches.

Due to the nucleophilic character of the Lys side chain, the residue is most commonly targeted with an electrophilic reagent. With a ε-NH_2 pKa of between 9.3-9.5, successful Lys conjugation is usually performed in basic conditions, to increase the proportion of the uncharged and nucleophilic amine. Selecting chemoselective reaction conditions is complicated however by the presence of competitive nucleophiles in the protein chain. In particular, the nucleophilic Cys side chain presents a competitive reactive site for many Lys labelling reagents, especially at near-neutral pH where the Cys residue is generally more reactive than Lys.[26]

A commonly used amine-labelling electrophile is found in the isothiocyanate group (Figure 1-11), which can be prepared from the reaction of the corresponding aniline with either thiophosgene, or the significantly less hazardous 1,1'-thiocarbonyldiimidazole (TCDI).[42, 43] The most prominent example of an isothiocyanate labelling reagent is fluorescein isothiocyanate (FITC), which is most often used as a fluorescent labelling reagent in a stochastic labelling approach. FITC is known to show cross-reactivity with thiols and alcohols, however the resulting products are less stable than the thiourea formed by reaction with an amine.[35] In addition to the issue of cross-reactivity, FITC is often sold as a mixture of two regioisomers which are difficult to separate. While these do not differ in their photophysical attributes, the heterogeneity of the reagent further underlines its use only as a non-specific labelling reagent.



Labelling method	Reaction conditions	Product	Comments
Isothiocyanate	pH 8-9.5, rt		Some cross-reactivity with other nucleophiles
NHS-ester	R: H, SO_3Na pH 7-8, rt		Very stable product, conjugation possible at near-neutral pH
Reductive amination	$NaCNBH_3/Na(OAc)_3BH$ pH 6.5-8.5, rt		Relatively slow, best in dehydrating conditions
Sulphonyl Halides	X: Cl, F pH 8.5-9.5, 20 °C		Reagent unstable in basic conditions, cross-reactivity with other nucleophiles

Figure 1-11 *Common methods for the modification of Lys residues in a protein substrate.*

A similarly popular family of reagents used for Lys labelling is that of the activated esters, with *N*-hydroxysuccinimide (NHS) ester perhaps the most commonly used variant. NHS acts as a good leaving group, and its solubility may be increased by addition of a sulphonate group to the succinimidyl ring (Figure 1-11). Some cross-reactivity with alcohol-bearing amino acid side chains has been reported when using NHS-esters, however the esters produced by this reaction are prone to hydrolysis and therefore less stable than the desired Lys labelled product.[33] While the amide bond formed with the Lys ε-NH_2 is extremely stable, the same cannot be said for the NHS-ester reagent, which will hydrolyse in aqueous basic conditions.[1] In a recent report, application of substoichiometric amounts of an NHS-reagent and careful choice of protein substrates have allowed residue-specific labelling of native sequence proteins.[44] Weil *et al.* have demonstrated this kinetically controlled labelling approach to the biotinylation of a protein by addition of 0.5 equivalents of the labelling reagent in 100 small volume portions.[44, 45] This work resulted in specific labelling of a single Lys residue in each protein explored, however, with an excess of protein substrate, the labelled product had to be

isolated by affinity chromatography before any future use. The explanation given for site-selectivity was preferential reaction at the most exposed Lys residue, therefore the specificity of the reaction depends on surface accessibility (this will be discussed further in Chapter 4).[44, 45]

Reductive amination is a synthetic strategy which is commonly used in Organic Chemistry for the production of secondary amines. Using a mild reducing agent such as $NaCNBH_3$ or $Na(OAc)_3BH$, the imine formed by condensation of a primary amine with an aldehyde is reduced to the corresponding secondary amine product. For use in protein-labelling, this reaction has primarily been applied to the labelling of proteins with carbohydrates, however dehydrating conditions are preferred to favour the initial condensation step and generally this approach has required long reaction times.[33] An additional drawback is the fact that some protein targets may contain imino groups which could undergo reduction with the borohydride reagent, such as in the porphyrin ring in haeme-containing proteins, or the chromophore structure formed within the GFP interior. However, Francis *et al.* have described the use of a water-tolerant iridium-based catalyst, which could be used to selectively reduce an imine formed by amine-aldehyde condensation, avoiding the use of the borohydride reagents.[46] Using a PEG-aldehyde, this system was found to label the target protein lysozyme with one or two PEG chains *via* reductive amination (lysozyme has six Lys residues in total).[46]

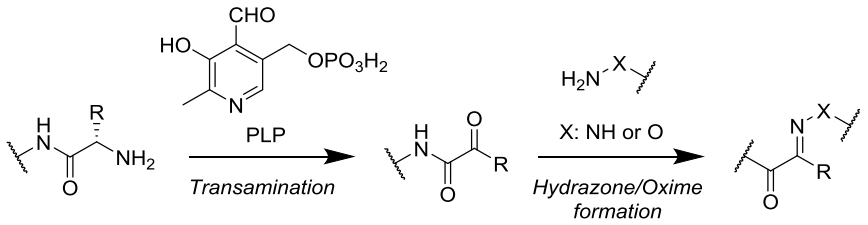
A final example of Lys modification comes in the form of sulphonyl halides, most frequently this involves the use of sulphonyl fluorides and sulphonyl chlorides. While the desired Lys-labelled product is a stable sulphonamide, sulphonyl halides also exhibit significant cross-reactivity with other nucleophilic residues. The reagent itself is also known to hydrolyse readily in basic conditions and therefore requires application in high excess compared to the substrate.[33] This being said, the compact nature of the $–SO_2X$ warhead has allowed genetic incorporation of a sulphonyl fluoride-functionalised UAA based on Tyr into the sequence of a target protein.[47] Applying this technology, Wang *et al.* observed covalent bridging within the modified protein, as well as cross-linking of interacting proteins *in vivo*, by covalent modification of native nucleophilic residues, including Lys.[47]

N- or C- terminal modification

The benefit of targeting the N- or C- terminus of a polypeptide is the potential for high site-selectivity. Unlike the various amino acid side chains, within a protein monomer there will exist only one N-terminus and one C-terminus. Of course, in proteins which exhibit quaternary structure, such as the cross-linked light and heavy chains in antibodies, there will be N-/C-

termini for every monomeric chain, however this is still a reduction in the number of reactive sites when compared with many amino acid targets. Due to their position at the beginning and end of the polypeptide chain it is also rare to find these reactive sites playing a role in the active site of a functional biomolecule, while accessibility is generally high.[48]

In terms of chemoselectivity, the N-terminus offers potential due to its low pKa (7.6-8) when compared with the Lys ε-NH_2 (pKa: 9.3-9.5), and this arises due to the proximity to the electron-withdrawing α-amide. In order to exploit this reactivity, many of the Lys-targeting chemistries which have been discussed previously may be applied to labelling of the N-terminus, with the adaptation that they are performed at neutral or weakly acidic pH where there will be a greater proportion of the uncharged, nucleophilic N-terminal α-NH_2 than there will of the Lys side chain. This being said, it should be noted that the pKa of the N-terminus will vary depending on the electron donating/withdrawing properties of the N-terminal amino acid side chain.[26] In addition it should be considered that as many as 57% of proteins in yeast and 84% of proteins in humans have been shown to exhibit at least partial acetylation of the N-terminal α-NH_2 *in vivo*, so in these cases targeting the N-terminus may not be possible.[49]

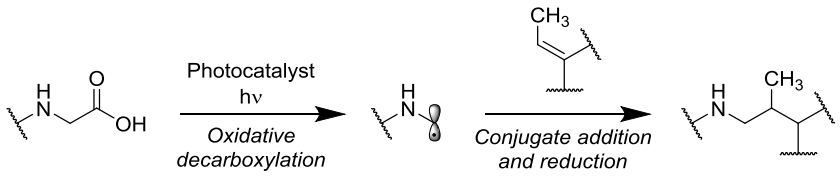


Scheme 1-1 *An example of an N-terminal-specific protein modification strategy. Using pyridoxal-5-phosphate (PLP) or a structurally related aldehyde-bearing pyridine, the α-NH_2 of an N-terminal residue may be converted to an aldehyde or ketone by transamination. This is converted to a hydrazone or an oxime through reaction with a labelled hydrazine or hydroxylamine.*

An interesting example of N-terminus labelling is highlighted above, where Francis *et al.* have used the natural cofactor pyridoxal-5-phosphate (PLP) to selectively convert the N-terminal α-NH_2 to a ketone or aldehyde by transamination (Scheme 1-1).[50] This process works by condensation of the amine with the aldehyde present in PLP, followed by imine-hydrolysis to yield the aldehyde (if the N-terminal amino acid is Gly) or ketone. Due to its proximity to the electron-withdrawing amide of the peptide backbone, the generated carbonyl species is prone to nucleophilic attack with a hydrazine or hydroxylamine to give the corresponding hydrazone or oxime. Using this system, Francis *et al.* have shown that it was possible to label proteins bearing Gly, Val, Lys and Met as the N-terminal amino acid, however there was poor conversion of substrates with His, Trp, Pro, Ser or Cys at this position.[50] Rather than relying

on a two-step labelling protocol, a PLP-related reagent which was based on a 2-pyridinecarboxaldehyde (2PCA) structure was developed by the same group which allowed labelling of the substrate with the label-bearing 2PCA reagent.[51] This reagent initially forms an imine with the N-terminal α-NH_2 as before, however it does not undergo imine hydrolysis and instead promotes cyclisation with the α-NH of the second amino acid in the chain. The resulting imidazolidinone structure was found to be reasonably stable, and this system was used to introduce labels such as fluorophores, biotin and PEG chains to a native sequence protein.[51]

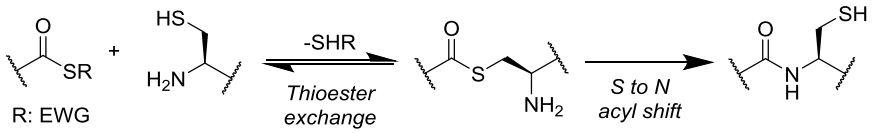
Compared to the published works describing N-terminal functionalisation, there is a comparative dearth of methods which selectively label the C-terminal α-CO_2H. This may hinge on the fact that there are few methods to distinguish between the reactivity of the C-terminus and the carboxylic acids contained on the side chains of Asp and Glu. However, just such a distinction was recognised by MacMillan *et al.*, who took advantage of the difference in oxidation potential of these sites, to selectively label the C-terminus in a photo-initiated decarboxylative alkylation reaction (Scheme 1-2).[52]



Scheme 1-2 *The C-terminal targeting photo-initiated decarboxylative alkylation reaction. Using an iridium or flavin-based photocatalyst, the C-terminus undergoes single electron oxidation, causing decarboxylation and generating the α-amino radical. Addition of a Michael acceptor results in conjugate addition of the radical species to the soft-electrophile, and the resulting radical species is reduced by single electron transfer from the photocatalyst.*[52]

Using an iridium or flavin-based photocatalyst under irradiation with blue LEDs, single electron oxidation of the C-terminus was achieved, followed by decarboxylation to give a reactive α-amino radical. Conjugate addition of the resultant radical to a Michael acceptor, and subsequent single electron reduction from the photocatalyst then gave the stable conjugate which had essentially been alkylated at the decarboxylated α-position of the C-terminal amino acid.[52]

While not explicitly a protein labelling method, native chemical ligation (NCL) is a technique used to covalently attach two peptide sequences in the synthesis of proteins, and represents a method which combines N- and C-terminal modification (Scheme 1-3).



Scheme 1-3 *An example of native chemical ligation (NCL), which allows peptide fragments to be stitched together for the synthesis of full length proteins. The C-terminal α-CO_2H is first activated as a thioester which generally bears an electron-withdrawing group (EWG) to make a good leaving group. A peptide with an N-terminal cysteine residue is introduced and reversible thioester exchange may occur. Through S- to N- acyl shift, the ligated product is then formed irreversibly.*

This approach requires a thioester-activated C-terminus and a Cys residue at the N-terminus, which undergo reversible thioester exchange to link the two peptide sequences. The thioester linkage may then undergo irreversible S to N acyl shift to give the native amide link between the peptides, with the preservation of the Cys stereocentre. In order to mask the point of ligation, many variants of NCL have been described, which include post-ligation desulphurisation to give an alanine residue, or the traceless Staudinger ligation, which does not require a Cys residue at the N-terminal.[53]

Tyrosine

Tyrosine is the target of choice for the work contained in this thesis, and it is an appealing target not only for its low abundance (~3% in the UniProt/EMBL database), but also due to its central position on the Eisenberg hydrophobicity scale.[29, 31] With its polar headgroup and aromatic ring, Tyr is considered amphiphilic and therefore can be found on the protein surface or in the protein interior. For this reason, it was considered an apt target for selective labelling, as in instances where there were multiple Tyr residues in a target protein, discrimination could be achieved based on surface accessibility.

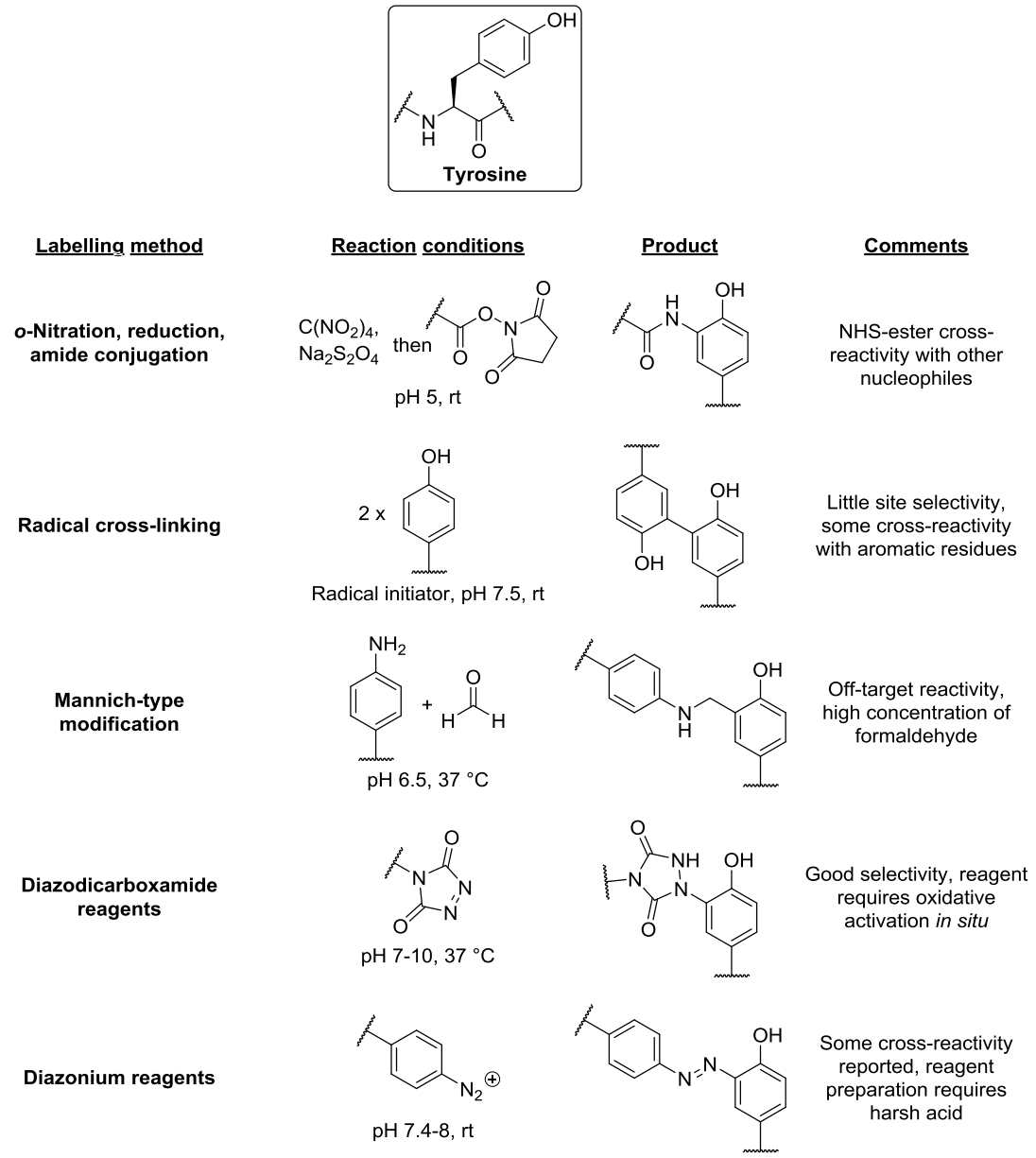


Figure 1-12 *Common methods for the modification of a Tyr residue in a target protein.*

Aside from the benefit of a low abundance and varied distribution in the protein structure, the Tyr side chain also offers nuanced reactivity through nucleophilic and radical-induced mechanisms (Figure 1-12). While there are several methods to modify Tyr residues reported in the literature, it has not proven as popular a target as Lys and Cys residues, however some key examples are highlighted below.

The first example relies upon nitration of Tyr using the nitrating agent tetranitromethane (TNM). In this approach, the Tyr phenol side chain is nitrated at the *ortho* position to give 3-nitrotyrosine, which is selectively reduced using a mild reducing agent such as sodium

dithionite. This reduction step results in formation of an *o*-aminophenol group, which may be modified by an NHS-ester at acidic pH.[54] Due to the low pKa of the aromatic amine (pKa ~5 for the ammonium form), it should be possible to selectively label the *o*-amino phenol at pH 5-7, where the Lys side chain and N-terminus are likely to be protonated and not nucleophilic.[33] However, in the work describing this method, competitive nucleophiles were acetyl-capped prior to reduction of the nitro group, suggesting selective modification of the modified Tyr was not easily achieved.[54] In addition, the reported cross-reactivity of the nitrating reagent TNM with residues including His, Trp and Met, suggest that this method is not well suited to use in selective protein labelling applications.[55]

Due to the electron-rich aromatic ring on the Tyr side chain, it is capable of single electron oxidation to the tyrosyl radical, which may undergo cross linking with a nearby aromatic or nucleophilic residue.[56] While this modification strategy has not been reported for protein labelling applications, it has been used as a means of affinity cross-linking, in which an associating protein is cross-linked to its target through radical-initiated cross-linking of Tyr residues.[56] In addition to Tyr-Tyr cross-linking, cross-reactivity with Cys, His and Trp have also been reported.[57]

In a more conventional protein-labelling approach, Francis *et al.* have described a three-component Mannich-type reaction, which was shown to label a Tyr residue at the electron-rich *ortho* position.[58] This approach relies upon *in situ* formation of an imine by an amine-bearing label and an aldehyde such as formaldehyde. This soft electrophile is subject to nucleophilic attack from the *ortho* position of the Tyr phenol ring, which is guided by a H-bonding interaction between the alcohol group and the nitrogen of the imine.[58] The resulting secondary amine is relatively stable, and this method was used to prepare a range of fluorescently labelled proteins, although in the example of protein substrate chymotrypsin, this resulted in labelling of two out of three native Tyr residues. Some cross-reactivity was highlighted with Trp residues and free Cys residues, but otherwise, this approach seems to provide good chemoselectivity for a Tyr target.[59]

Triazolinedione reagents are perhaps the most popular range of Tyr-labelling groups in current use. The disclosure of 4-phenyl-3*H*-1,2,4-triazoline-3,5(4*H*)-dione (PTAD) by Barbas *et al.* showed the modification of the phenol ring of the Tyr side chain *via* an ene-type reaction, which resulted in conjugation at the *ortho* position.[60] While some cross-reactivity with Lys and Trp residues was identified in single amino acid experiments, only Tyr modification resulted from a series of competition experiments.[60] Further expansion of the PTAD-reagent

scope resulted in the attachment of fluorophores, PEG groups and click handles to a range of Tyr-containing peptides and proteins with such high chemoselectivity that the bioconjugation reaction was referred to as the 'tyrosine click reaction'.[61] In aqueous conditions, the PTAD-derived reagents are prone to hydrolysis and for this reason, they are generally formed *in situ* by oxidation using *N*-bromosuccinimide and pyridine.[61, 62]

The final example given is the modification of Tyr residues by electrophilic aromatic substitution (S_EAr) using diazonium reagents. As with the PTAD example, this labelling reaction proceeds through the nucleophilic behaviour of the *ortho* position of the Tyr phenol ring. This approach has been used by Francis *et al.*, Barbas *et al.* and Haddleton *et al.* to selectively label Tyr residues in peptides and proteins.[63-65] The Francis and Barbas groups used diazonium reagents to install reactive ketone and aldehyde groups to biomolecular substrates through reaction with a Tyr residue.[63, 64] These tags were then used to attach functional labels through oxime-forming chemistry, resulting in selectively labelled biomolecules bearing fluorescent groups, biotin and PEG labels. In terms of selectivity, minor modification of His and Trp residues was observed in competition experiments with Tyr, however at much lower levels than the intended residue. Little or no cross-reactivity was observed with Cys, Ser or Lys.[64] A major benefit of this labelling approach is the production of an azobenzene group, which is labile to reductive cleavage.[66] This allows reversible coupling and uncoupling of a label to the Tyr target, while the amino acid, once cleaved, bears an *o*-aminophenol as evidence of modification. A drawback of using this chemistry in solution is the requirement to prepare diazonium salts in acidic conditions prior to application, as they tend not to be bench-stable reagents.

1.5 Conclusion

In this Introduction, a range of biomolecular substrates, labels and methods for protein modification have been explored. The array of available protein labelling methods is colloquially referred to as a 'toolbox', and within this set there remains a space for the development of an alternative strategy for the selective modification of a native sequence substrate. Given the low abundance, moderate surface-accessibility and nuanced reactivity of the Tyr side chain, this residue has been selected as the ideal candidate for the current work. The goal of this project is to provide a useful method for future protein labelling experiments, and in this way it is hoped that this work may expand the scope of the protein labelling toolbox.

1.6 Aims for this project

Following the assessment of the previously discussed strategies to achieve selective modification of a native sequence protein, the current work aims to develop a novel bioconjugation method which targets an existing Tyr residue in the protein substrate. Employing S_EAr chemistry, this work will also involve incorporation of a solid-phase support to provide superior selectivity for highly exposed Tyr residues over those which are buried in the protein interior. The successful development of this chemistry will allow the 'catch' and 'release' of a protein substrate from a solid-phase support, with the released protein bearing a novel modification which can be exploited for selective functionalisation. A range of conditions will be explored for the development of the 'catch' and 'release' steps of this reaction, and a suitable solid-phase support will be selected. Various approaches to the downstream functionalisation of the catch-and-release modified substrate will be investigated, and the preferred system will be applied to small molecule, peptide and finally protein substrates. The successful outcome of this work will be to deliver a three-step modification protocol which can provide selective functionalisation of a small protein with a synthetic label such as a fluorophore.

Chapter 2 Results & Discussion 1: Modification of a tyrosine residue

2.1 Introduction: Three-step modification of tyrosine

In previous work in the Hulme group, the utility of an azobenzene moiety as a cleavable linker has been explored.[66] In that application, the azobenzene group was installed in the early stages of the synthesis and it was employed purely as a cleavage point to be severed following reaction at the terminal end of the molecule. In the current work, azobenzene formation is achieved through electrophilic aromatic substitution (S_EAr) of the phenol ring of a Tyr residue, allowing the immobilisation of a biomolecule onto a solid-phase. Formation of the azobenzene moiety and subsequent reductive cleavage using dithionite therefore constitute the 'catch' and 'release' stages of a novel three-step protein modification strategy (Figure 2-1).

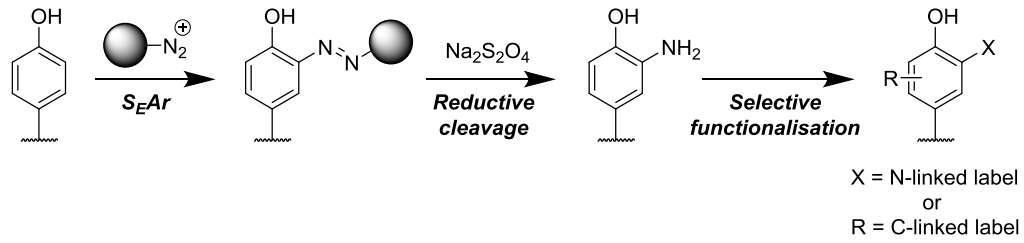


Figure 2-1 *The three-step strategy for modification of a Tyr residue.* ***S_EAr****: the 'catch' step involves immobilisation of the substrate through reaction of the Tyr phenol ring with a resin-bound diazonium salt to form an azobenzene moiety.* ***Reductive cleavage****: Using sodium dithionite, the azobenzene moiety is cleaved to release an o-aminophenol modified Tyr residue.* ***Selective functionalisation****: the unique o-aminophenol motif is exploited in a selective manner to equip the biomolecule with a functional label.*

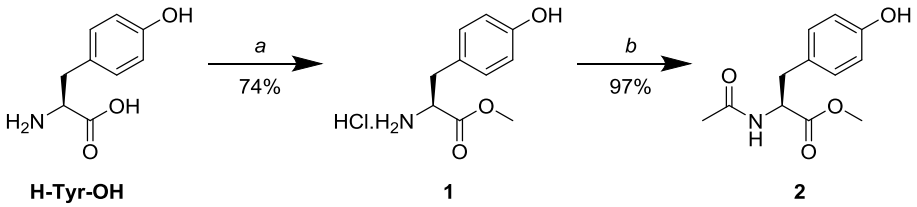
Due to its electron-rich phenol side chain, the Tyr residue can be modified through an S_EAr reaction with an electrophilic diazonium salt, forming the azobenzene species. Using a diazonium salt which is immobilised on a bulky solid-phase support provides a degree of selectivity in instances where multiple Tyr residues are present, as only the most exposed residues are available for modification. Following removal of any unbound substrate by washing the resin, the modified biomolecule is then released by introduction of the biocompatible reducing agent sodium dithionite.[66] Cleavage of the diazo group results in the release of an *o*-aminophenol modified substrate, which is then washed from the resin. The modified biomolecule now contains a unique synthetic handle among the endogenous reactive sites of a native sequence protein, and there are several reported strategies to exploit this in a selective manner. The post-cleavage functionalisation step should ideally allow the incorporation of a range of functional labels in order to expand the possible application of this modification strategy to as wide a base as possible.

While the final iteration of this work will feature a solid-phase platform, this chapter will discuss the initial development and optimisation of each of the constituent steps of the modification strategy. These were first explored in-solution using small molecule and peptide equivalents of the final biomolecular targets.

2.2 Synthesis of Tyr-bearing substrates

2.2.1 Ac-Tyr-OMe

For the development of the Tyr modification chemistry, simple small molecule and peptide substrates were used before progression onto more complex biomolecules. The N- and C-terminally protected amino acid Ac-Tyr-OMe **2** was used for much of this work and the compound was synthesised in two steps from H-Tyr-OH (Scheme 2-1).



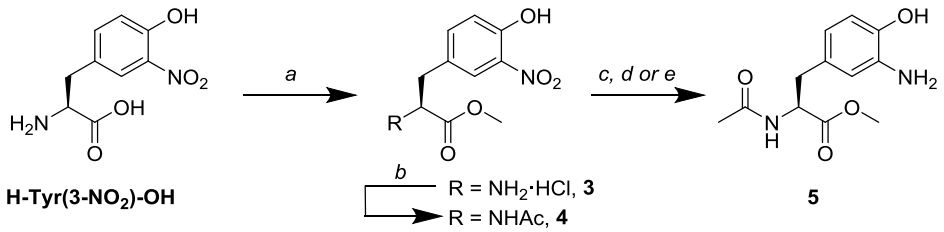
H-Tyr-OH **1** **2**

Scheme 2-1 *Synthesis of the N- and C-terminally protected Tyr residue **2**, which was used for the development of the Tyr modification chemistry. (a) MeOH, acetyl chloride, 0 °C – rt, 3 h; (b) Acetic anhydride, NaOAc, H_2O, 0 °C, 30 min.*

It was necessary to protect the reactive N- and C-termini to aid analysis of the Tyr-modification reactions by HPLC and TLC, as the highly polar H-Tyr-OH showed poor retention on a C_{18} column, and failed to move from the baseline on a normal-phase silica TLC plate. Initially, the carboxylic acid was converted to a methyl ester by Fischer esterification, with *in situ* generation of HCl by the reaction of acetyl chloride and MeOH to give **1**. Following purification by recrystallisation, the N-terminus was acetylated by addition of acetic anhydride in basic conditions. Performing the reaction in water resulted in the precipitation of the product **2**, which was washed with diethyl ether and isolated in high yield and purity.

2.2.2 Ac-Tyr(3-NH_2)-OMe

For development of the post-cleavage chemistry, an *o*-aminophenol modified Tyr residue was also synthesised. Commercially available 3-nitrotyrosine [H-Tyr(3-NO_2)-OH] was protected at the N- and C-termini in a similar manner to the H-Tyr-OH equivalent (Scheme 2-2). Reduction of the nitro group was then attempted, however this transformation proved difficult to achieve in high yield and purity and so several conditions were trialled to achieve clean conversion to the *o*-aminophenol species **5** (Table 2-1).



Scheme 2-2 *Synthesis of an o-aminophenol modified Tyr residue. Protection of the N- and C-termini of 3-nitrotyrosine followed the steps used for H-Tyr-OH. Various conditions were screened for reduction of the nitro group. (a) MeOH, acetyl chloride, 0 °C – rt, 3 h, 75h; (b) Acetic anhydride, NaOAc, H_2O, 0 °C, 30 min, 68%; (c) Fe, $CaCl_2$, EtOH/H_2O 95:5, 60 °C, 1 h, 62%; (d) 10% Pd/C, H_2, DCM, rt, 48 h, 23%; (e) $Na_2S_2O_4$, H_2O, rt, 90 min, 69%.*

Entry	Reducing agent	Workup and purification	Yield (%)	Notes
(c)[67]	Fe, $CaCl_2$	Passed through celite plug, flash chromatography	62	Heated to 60 °C, EtOH/H_2O solvent
(d)[68]	10% Pd/C, H_2	Passed through celite plug, flash chromatography	23	Low yield, long reaction time (48 h)
(e)[69]	$Na_2S_2O_4$	Salt precipitated with MeOH, flash chromatography	69	Reaction in water at rt, single reducing agent

Table 2-1 *Comparison of the three approaches taken to selective reduction of the 3-nitrotyrosine derivative. Purification of the o-aminophenol product by flash chromatography was complicated by its high polarity, resulting in lower than expected isolated yields.*

Initially, selective reduction of the nitro group was performed using Fe powder and $CaCl_2$; these conditions act through a catalytic transfer hydrogenation mechanism.[67] Purification of the resulting *o*-aminophenol was performed by flash chromatography, however this was complicated by the high polarity of the product, causing a narrow band on the silica gel to spread into a long streak as it progressed down the column. Nitro group reduction was also attempted using H_2 and 10% Pd/C; this reaction progressed slowly and ultimately resulted in a low yield of only 23%. The most operationally straightforward nitro reduction method involved the use of $Na_2S_2O_4$ in H_2O at room temperature. This reaction was judged to reach completion by TLC within 90 min, and the majority of the dithionite salt was removed by filtration after it was precipitated by addition of MeOH. Dithionite reduction was selected as the preferred method for this step due to its operational simplicity and production of the cleanest product; following flash chromatography a yield of 69% was achieved on a ~2 g scale.

2.2.3 Leu-Enkephalin

The pentamer peptide Leu-enkephalin (Figure 2-2) was selected as a substrate on which to demonstrate the Tyr modification strategy. Leu-enkephalin is an agonist of the δ-opioid receptor in humans and contains a single Tyr residue in its native sequence.[70, 71] It has been used in several literature reports to showcase Tyr-selective modification strategies, and so was considered appropriate for this work.[65, 72, 73]

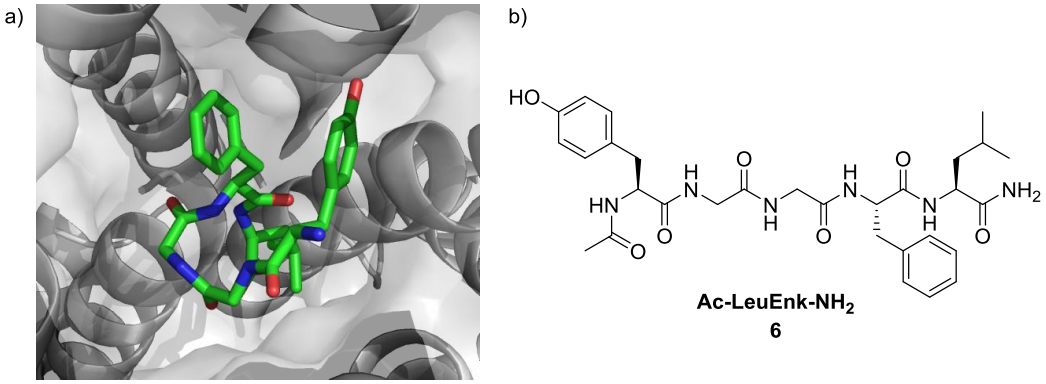


Figure 2-2 *Leu-enkephalin, the pentamer peptide chosen to demonstrate the Tyr modification strategy. (a) A model of the Leu-Enkephalin ligand bound to the human δ-opioid receptor. Structure retrieved from the PMDB database, PMDB ID: PM0079713.[71] (b) The structure of the synthesised Leu-enkephalin derivative, with an acetyl cap on the N-terminus and the C-terminus modified as a primary amide.*

Synthesis of the peptide was achieved by solid-phase peptide synthesis (SPPS) using Rink amide AM polystyrene resin and an Fmoc protection strategy (Figure 2-3). The resin is based on an aminomethyl polystyrene solid-phase, which must be swollen in DCM for maximum accessibility to the reactive sites. The resin bears a terminal Fmoc-protected amine and once deprotected in 20% piperidine in DMF, coupling of the first Fmoc-protected amino acid was performed using DIC and Oxyma coupling reagents. DIC was used due to the solubility of its urea by-product in organic solvents (in contrast to the insoluble by-product of DCC) which allows removal of the urea from the solid-phase by washing. Oxyma was included as a co-coupling reagent as this has been shown to reduce the occurrence of amino acid racemisation under coupling conditions.[74] Fmoc deprotection and amino acid coupling was repeated in an iterative manner until the desired peptide sequence had been achieved. After each deprotection or coupling step, a small portion of the beads were removed and tested for the presence of free amine using the 2,4,6-trinitrobenzenesulfonic acid (TNBS) test.[75] To perform this qualitative test, TNBS was added to the beads under basic conditions; under these conditions a free amine will attack the sulphonate-substituted position of the aromatic ring and displace the sulphonate leaving group, forming a red coloured conjugate. Any ambiguity in the TNBS test was

followed by repetition of the previous deprotection or coupling step in order to minimise the production of deletion-containing or truncated sequences, which make final purification of the desired peptide difficult.

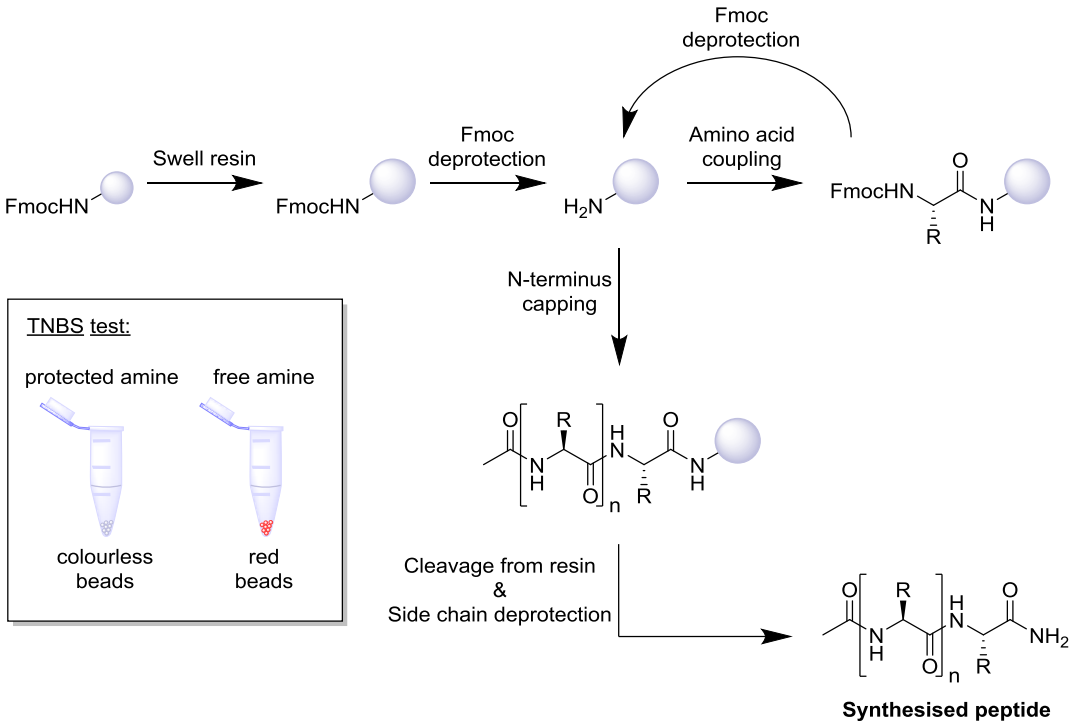
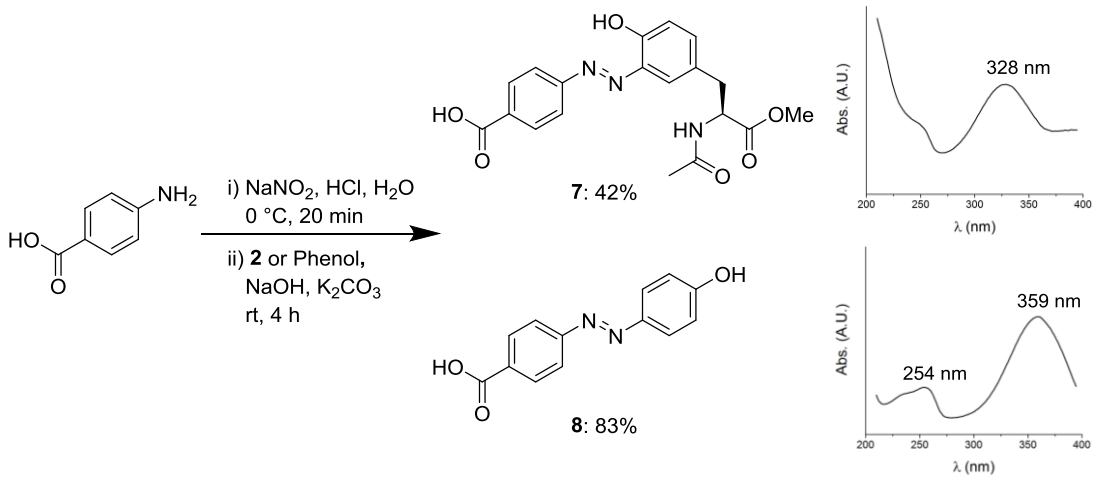


Figure 2-3 *The SPPS workflow. After swelling the resin, the Fmoc-protected terminal amine was deprotected. Amide coupling of an Fmoc-protected amino acid was performed, before deprotection of the newly installed amino acid. These two steps were reiterated until the desired peptide sequence was complete. The N-terminus was then capped with an acetyl group before cleavage from the resin and global deprotection of the amino acid side chains. After each coupling and deprotection step, a portion of resin beads were tested for the presence of free amine using the TNBS test (inset).*

After the desired sequence was complete, the N-terminus was capped with an acetyl group, before global deprotection of the acid-labile side chain protecting groups and cleavage from the resin. These steps were performed simultaneously by the introduction of a TFA cleavage cocktail which included triisopropylsilane (TIS) as a cation scavenger. Capping of the N-terminus as an acetamide and the C-terminus as an amide (which is achieved as a result of cleavage from the Rink amide AM polystyrene resin) was performed in order to simplify the synthesis, analysis and purification of the substrate. While this better mimics the behaviour of a mid-chain peptide, it should be noted that this also has a detrimental effect on the peptide's solubility. Analysis and purification of the synthesised peptide was performed by HPLC, giving the desired species Ac-LeuEnk-NH_2 **6** in 96% purity and 56% yield.

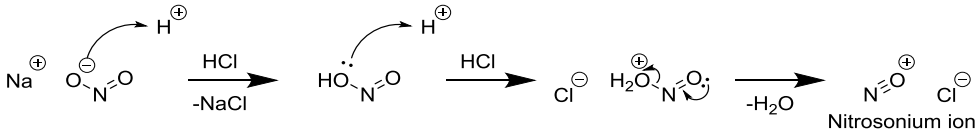
2.3 Electrophilic aromatic substitution

In the development of each step of the catch-and-release modification strategy, each reaction was tested on the simplified single amino acid substrate in-solution, before moving onto more complex substrates and the solid-phase. To demonstrate the S_EAr reaction with a phenolic substrate, *p*-aminobenzoic acid (PABA) was used to model an immobilised aniline, which would be used in the solid-phase system (Scheme 2-3), with the carboxylic acid group representing the anchor point to the resin. The aniline was first functionalised as a diazonium salt, before S_EAr with a phenol-containing substrate.



Scheme 2-3 *Azobenzene formation using Ac-Tyr-OMe* ***2*** *or Phenol as the nucleophile and PABA as the starting material for diazotisation. The percentages shown are isolated yields. The UV/Vis absorbance of each product is given (recorded by HPLC-PDA), showing the azobenzene absorbance at 328 and 359 nm respectively.*

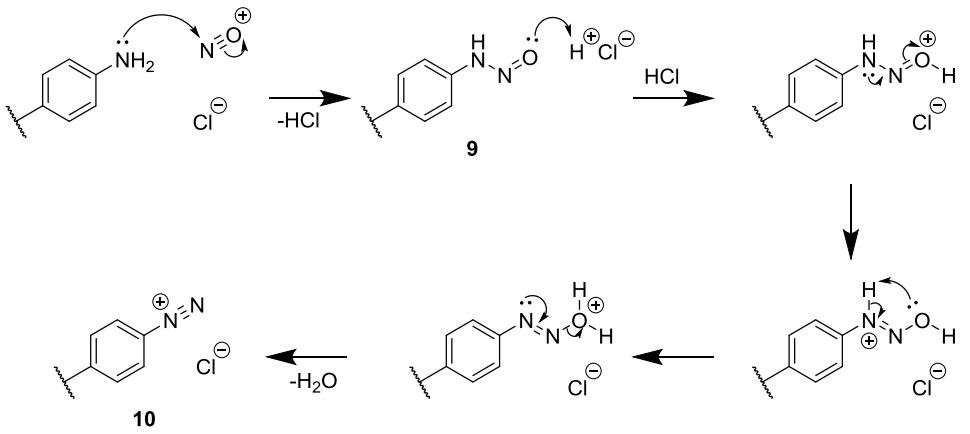
Diazotisation of PABA was achieved by addition of sodium nitrite under acidic conditions. These conditions allow *in situ* formation of the reactive nitrosonium ion (Scheme 2-4) by sequential protonation of the nitrite anion to form HONO and then H_2O^+NO. The nitrosonium ion is then formed by elimination of the H_2O leaving group.



Scheme 2-4 *Formation of the reactive nitrosonium ion from sodium nitrite in HCl.*

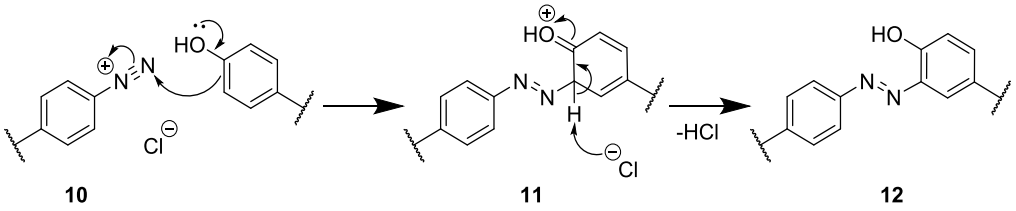
Nucleophilic attack of an aniline on the nitrosonium ion forms a nitrosamine intermediate **9** (Scheme 2-5). Sequential protonation of the terminal oxygen of the nitrosamine forms the

H_2O^+ leaving group, which allows formation of the diazonium salt **10** through the loss of a water molecule.



Scheme 2-5 *Diazotisation of an aniline by NO^+ in acidic conditions.*

While the diazotisation of an aniline by this method requires harsh acidic conditions, the following S_EAr reaction with a phenol proceeds at neutral or basic pH. Although the substitution reaction is accelerated in basic conditions, neutral pH was preferred in this instance for compatibility with future protein substrates. Therefore, after diazotisation the reaction mixture was neutralised to ~pH 7 using NaOH and K_2CO_3, monitored using pH paper. Introduction of a phenol at this stage resulted in azobenzene formation through the S_EAr mechanism (Scheme 2-6). The terminal nitrogen of the diazonium salt first undergoes nucleophilic attack from the phenol ring to give the intermediate **11**. Deprotonation of this intermediate restores aromaticity in the phenol ring and results in formation of the stable azobenzene **12**. Due to the absorbance at ~350 nm of the azobenzene moiety, the reaction mixture took on an orange/red colour as the product formed.



Scheme 2-6 *Electrophilic aromatic substitution of a para-substituted phenol with a diazonium salt, resulting in modification at the ortho position of the phenol ring. In Tyr, the para position is blocked by the β-carbon of the amino acid side chain.*

Due to the *ortho*- and *para*- directing effects of the OH group, it was expected that we would achieve *ortho*-substitution of the Ac-Tyr-OMe substrate upon reaction with a diazonium salt, as the *para* position is blocked by the β-carbon of the amino acid. By NMR, the azobenzene

product from this reaction **13** exhibited the expected splitting pattern corresponding to single modification of the phenol ring at the *ortho* position (Figure 2-4a). By comparison, the splitting pattern observed upon azobenzene modification of phenol suggested substitution at the *para* position, resulting in a plane of symmetry in each aromatic ring of azobenzene product **14**. Despite the *ortho*- and *para*- directing effects of the OH in phenol, no *o*-substituted product was observed in this case, perhaps because the *para* position is less sterically hindered than the *ortho* positions.

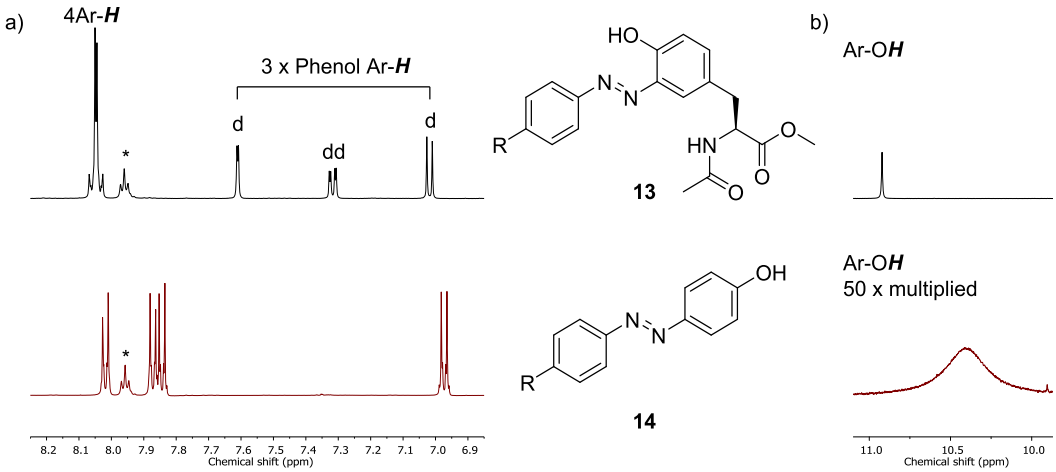


Figure 2-4 *Ortho- and para- substitution of different phenolic substrates. (a) The Tyr-derived azobenzene (top) showed ortho-substitution, as demonstrated by the d, dd, d splitting patterns of the aromatic peaks between 7.0 and 7.7 ppm. The phenol-derived azobenzene (bottom) demonstrates para substitution, resulting in symmetrical peaks in the aromatic region which each integrate to 2 protons. The peak heights were normalised to an amide peak (*) present in the R group. (b) The OH of the Tyr-derived azobenzene gave a surprisingly sharp peak, presumably due to H-bonding with a nitrogen in the diazo group. The phenol OH in the para-substituted compound was extremely broad by comparison. Both NMR spectra were recorded in DMSO-d_6.*

A recurring theme in the NMR of Tyr-derived azobenzenes was the observation of a surprisingly sharp OH peak, which was not seen in the *para*-substituted phenol-derived products (Figure 2-4b). This was justified by potential H-bonding between the OH of the *ortho*-substituted phenol and a nitrogen of the diazo group, which would stabilise the normally labile proton and reduce peak broadening. This postulation was supported by the observation of a very broad OH peak in *para*-substituted azobenzenes derived from phenol.

With a synthetic route to the azobenzene-linked Tyr now established, the chemistry of the 'catch' step was successfully demonstrated. The next step in the modification strategy was to cleave the azobenzene linker to release the modified Tyr-bearing cargo.

2.4 Reductive cleavage of the azobenzene moiety

The use of azobenzene as a cleavable motif for reversible bioconjugation was first described by Sharon *et al.*, who developed a bifunctionalised cross-linking agent which was used to cross-link membrane proteins.[76] At the core of the linker was an azobenzene moiety which allowed decoupling of the protein conjugates under the mild reducing conditions of aqueous sodium dithionite ($Na_2S_2O_4$). Biotinylated azobenzene-containing linkers have since been employed by Bogyo *et al.*,[77] Hulme *et al.*[66] and Wagner *et al.*[78] to enrich protein samples through exploitation of the interaction of biotin with its avidin or streptavidin binding partner. This approach allowed the researchers to exploit these famously strong binding interactions with subsequent release of the enriched cargo under mild conditions. Besides the dithionite cleavage strategy used in this work, various reducing agents have been reported to cleave the diazo bond which is central to the azobenzene group. Among these approaches, metal-mediated reactions employing titanium,[79] nickel,[80] zinc,[81] and palladium[82] have all been shown to achieve reductive cleavage of the diazo bond to give the corresponding anilines. Considering a biochemical approach, azoreductase enzymes found in the intestine have been shown to achieve the same transformation, however these are intracellular enzymes which are not particularly structurally stable, limiting the scope of their application in a reaction flask.[83] Another alternative reducing agent is found in the form of hydrazine hydrate, however, despite a standalone account of azobenzene cleavage using this reagent,[84] subsequent work demonstrated that this reagent was successful in reduction of the azobenzene to the hydrazine intermediate, but did not achieve complete cleavage to the respective anilines.[85-87] Compared to these alternatives, dithionite is an attractive reagent for azobenzene cleavage, as it is commercially available, inexpensive, water soluble and mild with respect to protein substrates.[83] In addition, using an excess of dithionite results in quantitative conversion to the fully reduced aniline products.[88] This is particularly important in the solid-phase system envisioned, as partial reduction to the hydrazine product would not allow release of the modified substrate from the resin.

Scheme 2-7 *The reductive cleavage of an azobenzene group is achieved through reduction to the intermediate hydrazine, followed by reduction to the corresponding anilines. Each step involves the addition of two electrons and two hydrogens to the azobenzene substrate.*

The mechanism of reductive cleavage of an azobenzene moiety relies on the delivery of a total of four electrons and four protons from an appropriate reducing agent (Scheme 2-7). This transformation can be seen as two sequential steps, with the initial reduction of the azobenzene group to its hydrazine equivalent; and following this, further reduction to the corresponding anilines. The current work uses the mild reducing agent sodium dithionite to sever the diazo double bond, and while the exact mechanism remains the subject of discussion, it is thought that the ability of dithionite to act as a reducing agent relies on its degradation in water to form two $SO_2^{-\bullet}$ molecules. This radical anion is able to act as a single electron reducing agent to donate the electrons necessary for diazo reduction, with protons coming from the aqueous solvent.[83]

Cleavage trials were undertaken using a Tyr-bearing azobenzene and different concentrations of aqueous dithionite (Figure 2-5a). The cleavage reaction was monitored by UHPLC-PDA, where the azobenzene starting material (with a distinctive absorbance at 330 nm) was converted to two anilines with shorter retention times and distinct PDA spectra. The conversion was also noticeable by eye, as a bright orange solution of the azobenzene species in 300 mM dithionite became completely colourless within 1 min (Figure 2-5b).

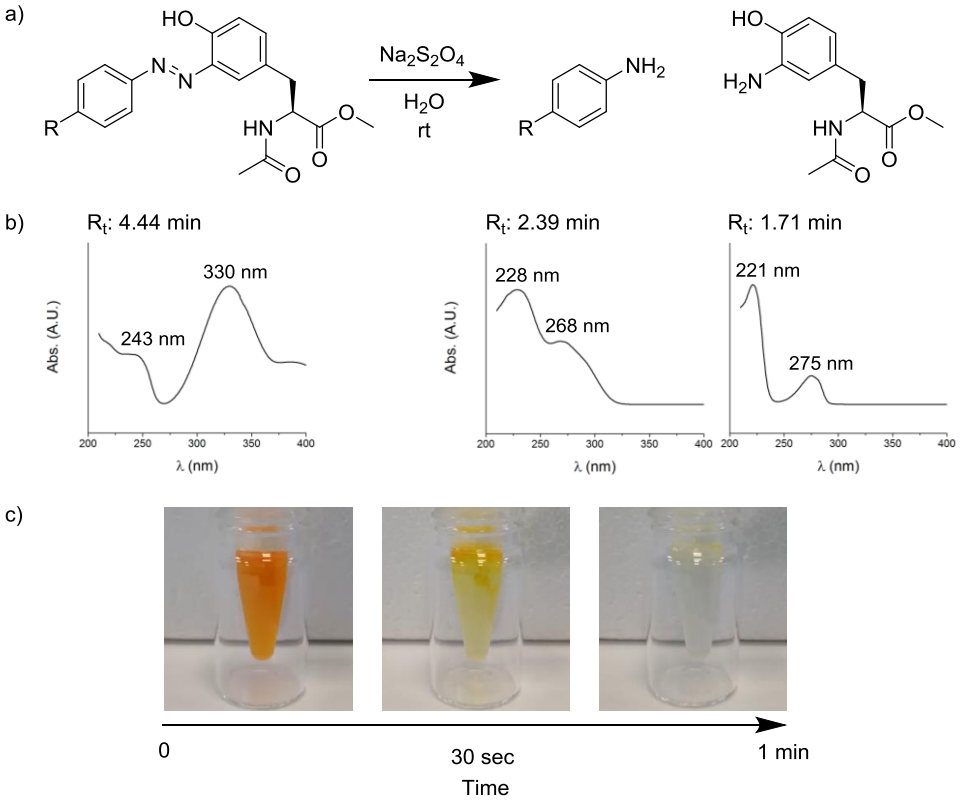


Figure 2-5 *Reductive cleavage of the azobenzene moiety in solution. (a) The reaction conditions for the azobenzene cleavage trials. (b) The azobenzene starting material has a distinctive absorbance at 329.5 nm, which is lost upon reductive cleavage, and not seen in either cleavage product (R_t shown is from UHPLC, Method 2, System 3). (c) Frames from a video of a solution of azobenzene in 300 mM dithionite which was completely colourless within 1 min.*

300 mM aqueous dithionite was the highest concentration used in the trials, however it was found that in solution, dithionite concentrations of 300 mM, 100 mM, 50 mM, 10 mM and 5 mM all resulted in complete conversion to the cleavage products by the first time point of 15 min. Within this time frame, none of the semi-reduced hydrazine derivative was observed. While this represents extremely efficient reduction of the azobenzene group, prior reports of the use of dithionite for reduction reactions on solid-phase call for concentrations up to 300 mM, and this was heeded for the future application of reductive cleavage on an immobilised azobenzene substrate.[66]

2.5 *o*-Aminophenol as a functional handle

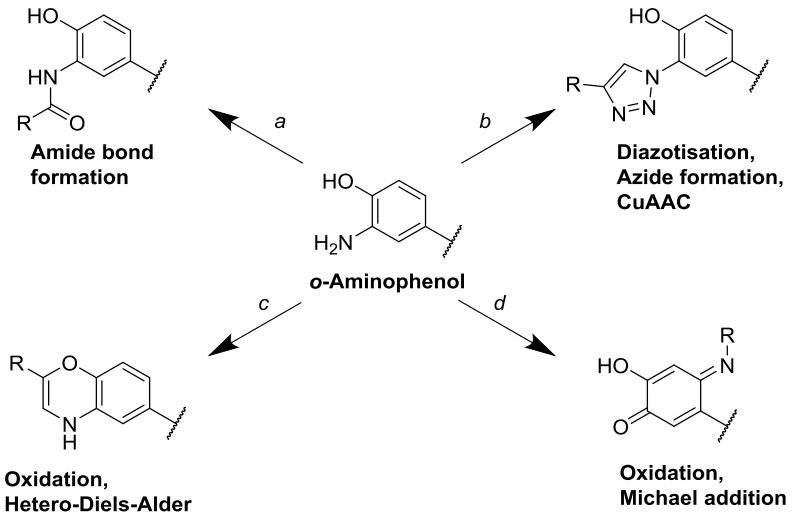
The catch-and-release protein modification system produces an *o*-aminophenol modified substrate, which must then be labelled with a functional molecule. Successful labelling of the catch-and-release modified protein requires selectivity for the *o*-aminophenol handle over other reactive sites, and this reaction should proceed under biocompatible conditions. The chosen modification strategy should therefore involve biocompatible reagents, it should be efficient at low concentrations and should occur in a primarily aqueous solvent, at a mild pH and temperature.[89]

While the *o*-aminophenol motif is considered to be unique in the proteins of interest for this study, it should be noted that there is potentially a biochemical route to the same product. Nitration of a Tyr residue to give 3-nitrotyrosine is a known PTM,[90] and subsequent reduction to 3-aminotyrosine may be possible by reaction with biological reducing agents (previously this has been demonstrated with haeme-containing proteins).[91, 92] The Tyr nitration PTM arises through exposure to reactive nitrogen-based species, but is not highly abundant and appears to show some selectivity for specific residues in certain proteins.[93, 94] If Tyr nitration is a concern in a protein with a known amino acid sequence, it can be readily identified by MS. Alternatively, where the sequence of the protein is unknown, Tyr nitration can be identified using an anti-nitrotyrosine antibody.[95*] The majority of literature reports which describe *o*-aminophenol-specific modification come from proteomic approaches to quantify Tyr nitration in protein samples, and therefore require nitro-reduction to furnish the reactive species. As described previously, dithionite has proven an efficient and biocompatible reducing agent for conversion of 3-nitrotyrosine to 3-aminotyrosine, and is therefore often used for this conversion.

The selected 3-aminotyrosine modification approaches shown in Scheme 2-8 can be grouped based on the manner in which they target the *o*-aminophenol moiety. One approach is to exploit the aromatic amine which has been installed on the Tyr ring. This amine has a lower pKa than the aliphatic amines present in the protein sequence and so, theoretically, can be selectively modified in an amide bond forming reaction, or in a route requiring diazonium salt formation (Scheme 2-8a & b). An alternative approach is to exploit the redox potential of the electron-rich aromatic ring which now bears two electron donating substituents in OH and NH_2. By oxidation of this system using a biocompatible oxidant, the resulting *o*-iminoquinone

[*] All proteins used for this project were commercially sourced and no PTMs were identified by the vendors. The most frequently used proteins were analysed by MS and the observed mass matched the theoretical mass of the native sequence.

may be modified using hetero-Diels-Alder or Michael addition chemistry (Scheme 2-8c & d). Each of these approaches were explored in more detail in order to select the most appropriate post-cleavage modification method.



Scheme 2-8 *Selected approaches to the functionalisation of an o-aminophenol moiety. (a) Amide bond formation with the aromatic amine at pH 5. (b) Conversion of the aromatic amine to an azide, before CuAAC with an alkyne (c) Oxidation to form a diene, before hetero-Diels-Alder reaction with a dienophile. (d) Oxidative coupling by Michael addition using an oxidising agent and a nucleophilic aniline.*

2.5.1 Amide bond formation

In the area of proteomics, quantification of Tyr nitration in various disease models has been investigated as a potential marker for oxidative and nitrosative stress.[96] While detection of 3-nitrotyrosine using antibodies is well established,[95] MS techniques have been used to identify the specific sites of modification in nitrated proteins. Unfortunately, nitro modification of a Tyr residue appears to limit peptide fragmentation by electron-based techniques, and so to circumvent this, several derivitisation methods have been reported. Nitro reduction to give the *o*-aminophenol group has been shown to increase peptide coverage by electron-based fragmentation techniques,[97] and specific labelling of this group has also been used to attach small organic labels (including biotin) to identify and enrich the previously nitrated species.[90] As the only aniline in an otherwise unmodified biomolecule, the aromatic amine in *o*-aminophenol has a lower pKa (4.8 in its protonated form) than aliphatic amines which could act as competitive nucleophiles.[98] Therefore, with careful pH control it is theoretically possible to favour modification of this site over other endogenous nucleophiles (Figure 2-6). Most

commonly, *o*-aminophenol modification has been achieved through coupling to an NHS-ester,[54, 94, 99] however a similar approach has been reported with a sulfonyl chloride.[96]

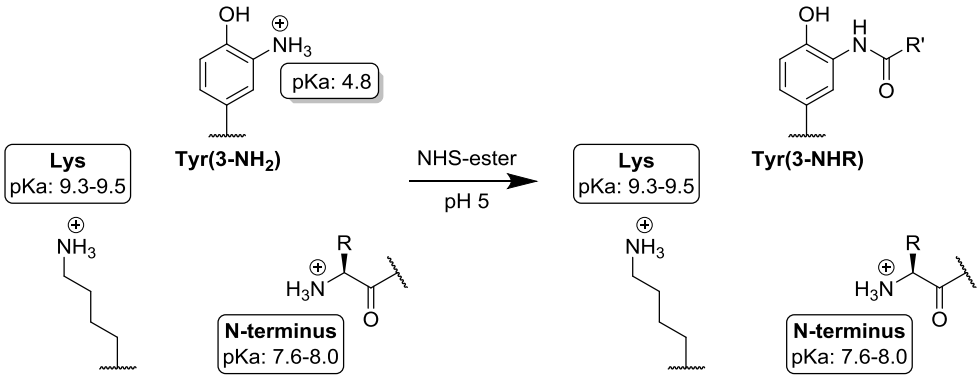
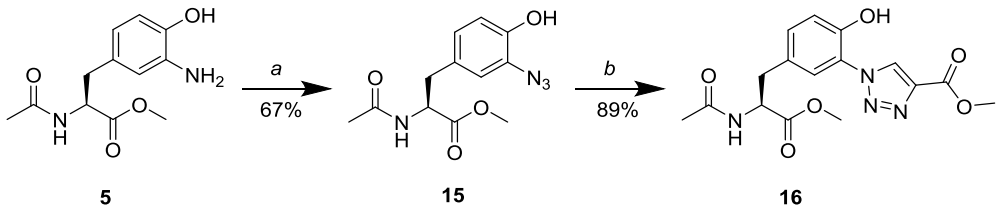


Figure 2-6 *Selective modification of o-aminophenol using an NHS-ester labelling reagent. Due to the lower pKa of the aromatic amine compared with the aliphatic amines such as the Lys side chain and the N-terminus, selective modification should be possible with careful pH control.*[48, 54]

At pH 5, the aniline of *o*-aminotyrosine is mostly deprotonated and behaves as a nucleophile, while the Lys side chain and N-terminus remain mostly in their charged forms. However, these pKa values are an average descriptor of each residue, and electrostatic effects in the local environment, as well as the degree of solvation of the residue, can have a significant impact on their effective pKa.[100, 101] For example, the ε-NH_2 of a Lys residue which is surrounded by basic residues is likely to have a lower pKa and therefore appear more nucleophilic than a residue in an acidic local environment. Likewise, the pKa of the α-NH_2 of the N-terminus varies depending on the electron withdrawing or donating properties of the amino acid sidechain. In addition, the thiol sidechain of Cys is a potentially competitive nucleophile. Perhaps for these reasons, amide bond forming reactions of *o*-aminophenol modified residues have not yet been reported on intact proteins, and instead have been used only on enzymatically digested peptide fragments. In these examples, amines in the peptides which might react in competition with the *o*-aminophenol were acetylated while the Tyr modification was still in its nitro- form, before reduction and amide bond formation with the single free amine.[94, 99] This would suggest that the reaction with an NHS-ester at pH 5 is not sufficiently selective to avoid off-target modification of other endogenous nucleophiles, and so this approach was not considered to be suitable for this project.

2.5.2 Conversion to an azide and CuAAC

The *o*-aminophenol moiety enjoys a privileged position, bearing the only aromatic amine in an otherwise unmodified native sequence protein. Using now familiar chemistry it is therefore possible to diazotise this group to facilitate selective labelling (Scheme 2-9). Through treatment with sodium nitrite in acid, the aromatic amine readily forms a diazonium salt. While in the S_EAr reaction, treatment of a phenol with this diazonium species would result in azobenzene formation; using sodium azide as the nucleophile, it is possible to convert the diazo group directly to an aromatic azide (as in **15**), which can be selectively modified using copper-catalysed azide-alkyne cycloaddition (CuAAC).[102] This methodology was tested using the commercially available alkyne methyl propiolate, which resulted in the triazole-functionalised Tyr residue **16**.

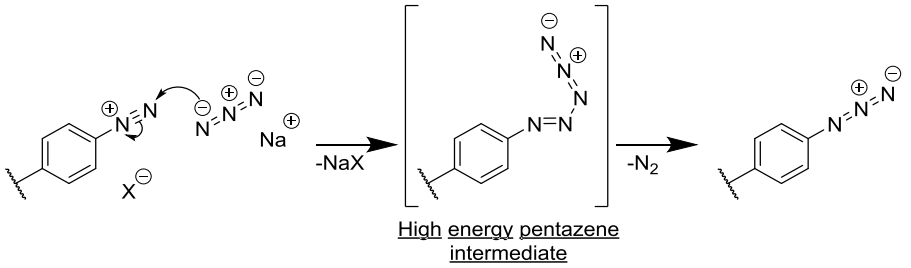


Scheme 2-9 *Conversion of an o-aminophenol to an aromatic azide, followed by CuAAC. (a) i) HCl, $NaNO_2$, H_2O, 0 °C, 10 min. ii) NaN_3, H_2O, 0 °C, 6 h. (b) Methyl propiolate, $CuSO_4 \cdot 5H_2O$, sodium ascorbate, t-amyl alcohol/H_2O, rt, 16 h.*

It is important to note the potential dangers associated with the azide formation reaction, as it involves addition of sodium azide to a solution which had previously contained HCl. Under acidic conditions sodium azide is known to form the toxic, volatile and explosive hydrazoic acid, and so this reaction was performed behind a blast shield and only on a small scale (<1 mmol).

The generic substitution of an aromatic diazonium salt with a nucleophile is a well-studied transformation. A notable example is the Sandmeyer reaction, which involves the functional group transformation of a diazonium salt to a halide. Adaptations of the Sandmeyer reaction include the use of thiol, alcohol and cyano nucleophiles among others, while recent reports apply similar conditions to install a trifluoromethyl group or a radiolabelled iodine onto a diazotised substrate.[103, 104] The use of an azide anion as a nucleophile in a Sandmeyer-type transformation has also been well explored due to the utility of the azide group as a masked amine or in various 'click' chemistries.[102, 105-107] Despite this fact, the mechanism by which this transformation occurs has only recently been described.[108]

Using a combination of theoretical and experimental methods, Cossio *et al*. found that, contrary to the previously proposed S_N2 mechanism, the conversion of a diazonium salt to an azide is in fact instigated by addition of the azide anion to the terminal nitrogen of the diazonium salt (Scheme 2-10). It was found that a high energy pentazene intermediate was formed, which then loses N_2 to give the desired aromatic azide.[108] An interesting implication of this revised mechanism is that only the terminal nitrogen in the azide product has actually originated from the azide nucleophile, while the other two nitrogens were already present in the diazonium salt. Another point of interest is the similarity between the initial step of this mechanism and that of the S_EAr reaction used to form an azobenzene. The difference in the case of azide formation is that the pentazene intermediate formed by the initial nucleophilic attack on the diazonium salt is very high in energy and therefore short-lived, while the azobenzene product is sufficiently stable to be isolated and purified.

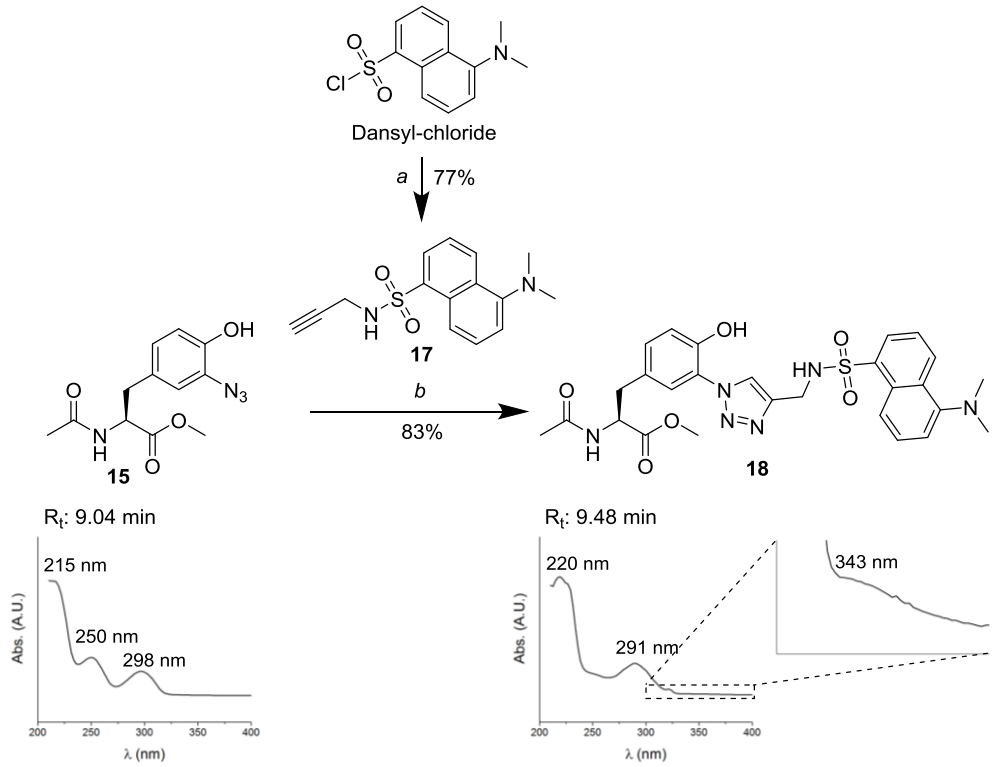


Scheme 2-10 *The azide conversion of a diazonium salt. Contrary to the expected substitution reaction, the azide-forming reaction is now thought to progress via addition of the azide to the diazonium salt to form an unstable pentazene intermediate.[108] Following the loss of N_2, the aromatic azide is formed.*

Following successful functionalisation with an azide, a number of bioorthogonal reactions were available for the modified Tyr residue. The preferred approach in this instance was to use CuAAC, due to the relative ease with which alkyne-functionalised fluorescent labels are synthesised, the aqueous compatibility of the system and the low cost of the Cu catalyst.[109] Alternative azide labelling reactions include Ru-catalysed azide-alkyne cycloaddition (RuAAC)[110], strain-promoted azide-alkyne cycloaddition (SPAAC)[111], and Staudinger ligation.[112]

A fluorescent alkyne in the form of dansyl alkyne **17** was synthesised in high yield from the commercially-available dansyl chloride. This was added to the azido-tyrosine derivative **15**, and the Cu^+ species was generated *in situ* in a *t*-amyl alcohol/H_2O solvent mix. The reaction was monitored by HPLC-PDA, and in addition to the distinct retention times of the starting materials and product, the PDA absorbance spectra provided supplementary evidence of the functional group conversion. The azido-Tyr derivative **15** exhibited a distinctive absorbance

at ~296 nm, which has previously been shown to arise from azidophenol compounds.[113] In the CuAAC product, the absorbance of the dansyl group was largely swamped by that of the Tyr moiety, however a shallow absorbance at around 340 nm was also observed, and this is supported by literature reports of similar triazole-linked dansyl groups (literature example λ_{ex} 335 nm, λ_{em} 520 nm).[114] The purified click product **18** was further characterised by NMR and MS, confirming its identity, but the HPLC-PDA method proved a useful tool to monitor the reaction.

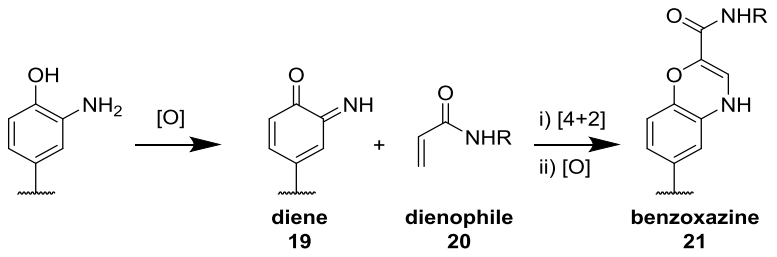


Scheme 2-11 *Synthesis of a fluorescent Tyr derivative using CuAAC chemistry. (a) Propargylamine, NEt_3, DCM, 0 °C, 1 h. (b) $CuSO_4 \cdot 5H_2O$, NaAsc, t-amyl alcohol/H_2O (2:1), rt, 18 h.*

A high yield of the click-conjugated Tyr derivative **18** was achieved in 18 h at room temperature, demonstrating a potential route to functionalisation of a catch-and-release modified Tyr substrate. However, there are some issues with the diazotisation, azide-formation CuAAC modification strategy, not least the need for harsh acidic conditions to diazotise the *o*-aminophenol moiety in the first step. As such, it would seem that this modification strategy is best suited to small molecule and peptide substrates, as protein substrates are likely to become denatured in these conditions. Further to this, in looking forward to a future protein substrate it would be preferable to minimise the number of reaction steps and reduce the need for purification wherever possible.

2.5.3 Hetero-Diels-Alder approach

Many bioconjugation strategies rely on the nucleophilic character of a target residue, and the nucleophilicity of the amine in the *o*-aminophenol moiety has so far been explored in the amide bond-forming and diazotisation-azide formation approaches. Selectivity based on nucleophilicity is complicated by the presence of competing nucleophiles among the 20 canonical amino acids. In particular, nucleophilic amines, thiols and hydroxyl groups pose potential competition for reaction with an electrophilic modifying agent. An alternative modification approach involves oxidative activation of the *o*-aminophenol moiety (Scheme 2-12). This approach opens the possibility for more nuanced reactivity with a complementary labelling reagent and therefore may achieve greater selectivity for the modified residue over potentially competitive sites.



Scheme 2-12 *Hetero-Diels-Alder modification of an o-aminophenol. When exposed to a suitable oxidising agent, an o-iminoquinone is formed. An appropriate amide-substituted dienophile bearing a functional synthetic group (R) allows formation of a morpholine ring via [4+2] cycloaddition. Subsequent oxidation results in the unsaturated benzoxazene* ***21****.*[115]

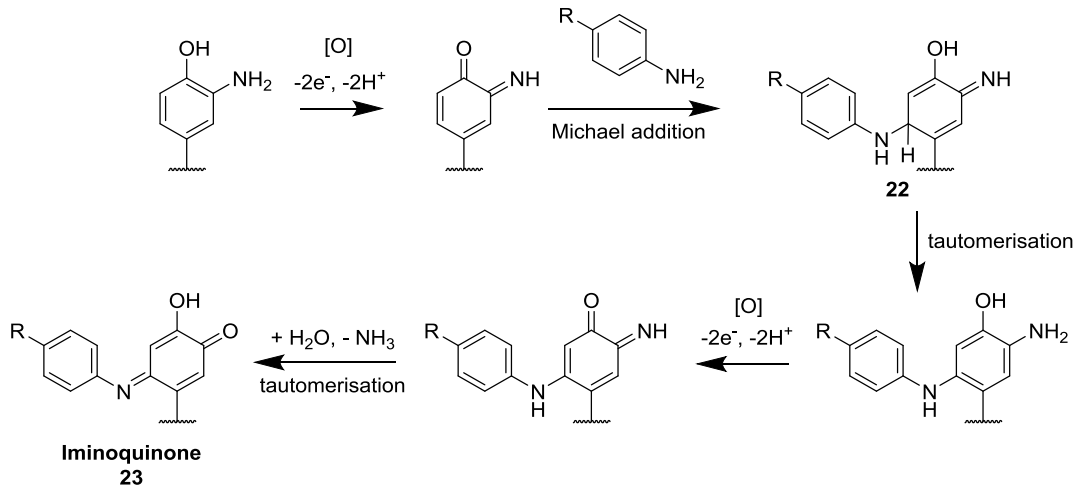
As one of the few electron-rich aromatic amino acids, Tyr is known to be redox active,[90] and oxidation of Tyr and its derivatives have been reported *via* both enzymatic and chemical routes.[57] Oxidation of a Tyr-related residue results in the production of a higher oxidation state species such as a quinone or iminoquinone. The tyrosinase enzyme has been reported to oxidise solvent-exposed Tyr residues, which can then participate in selective covalent modification.[116, 117] In terms of chemical approaches, Francis *et al*. have pioneered selective oxidation techniques to target Tyr derivatives with small molecule oxidising agents.[89] Initially, much of this work involved the use of sodium periodate as an oxidant of *o*-aminophenols. Other oxidising agents which have been explored include ceric ammonium nitrate and a range of silver, cerium and copper salts; however the most promising biocompatible oxidant which was assessed was potassium ferricyanide [$K_3Fe(CN)_6$].[89] The ferricyanide oxidant has since featured in several reports, where it has been used to activate a Tyr-derived target for bioconjugation. Crucially, this mild oxidant has shown efficient oxidation of catechol and *o*-aminophenol groups (to *o*-quinone and *o*-iminoquinone respectively), but does not oxidise off-target sites such as the side chains of Cys, Met and Trp residues.[89] A benefit of using a small

molecule oxidant compared to an enzymatic reagent is the ease of handling, as well as straightforward separation of the oxidising agent from the protein substrate following modification.

Upon oxidation of an *o*-aminophenol, the hydroxyl group is converted to the ketone form, and the amine to its imino counterpart. This *o*-iminoquinone (Scheme 2-9 **19**) has been used as a tuned diene for hetero-Diels-Alder mediated modification using a suitable dienophile in the form of an acrylamide (**20**).[115] The hetero-Diels-Alder reaction is an appealing method for bioconjugation as these reactions tend to progress with a high rate constant and are feasible in aqueous buffer at mild temperatures.[118] However, this reaction has also been shown to produce a heterogeneous mixture of conjugation products, which is undesirable as analysis and purification of protein bioconjugates is more complex than with their small molecule counterparts.[119, 120] In addition, attempts to recreate this transformation on small molecule substrates were unsuccessful in our hands, and for this reason the hetero-Diels-Alder strategy was abandoned in favour of an alternative oxidative coupling strategy.

2.5.4 Oxidative coupling

Following the report of the hetero-Diels-Alder strategy for modification of an *o*-aminophenol, Francis *et al.* developed an alternative oxidative coupling reaction which supersedes its predecessor in terms of selectivity.[89] Over several reports, this coupling strategy has been shown to exhibit high efficiency under biocompatible conditions, with no reported cross-reactivity or oxidation of susceptible native residues.[121-124] The bioconjugation product is achieved using the favoured $K_3Fe(CN)_6$ oxidant at pH 6.5. As before, oxidation of the *o*-aminophenol moiety to its *o*-iminoquinone equivalent activates the modified Tyr residue for selective bioconjugation (Scheme 2-13). In this instance however, addition of a nucleophilic aniline results in Michael addition to the modified Tyr side chain, giving the conjugate addition product **22**. Following further tautomerisation and oxidation steps, the imine group undergoes hydrolysis to afford the desired bioconjugate in the form of iminoquinone **23**. Through computational modelling and NMR studies, Frances *et al.* have shown that this is the preferred conformation of the oxidative coupling product,[121] and the iminoquinone product was also found to exhibit high stability across a range of temperatures, pH and upon incubation with various nucleophiles.[121]



Scheme 2-13 *The mechanism of the oxidative coupling reaction. Oxidation of the o-aminophenol forms the reactive o-iminoquinone structure. Conjugation of the label is achieved by Michael addition of the aniline to the o-iminoquinone. Following tautomerisation and further oxidation, the imine is hydrolysed to give the iminoquinone product.*

Initial attempts to recreate the oxidative coupling reaction involved the use of simplified substrates: 2-amino-*p*-cresol (to mimic the modified Tyr residue) and *p*-toluidine (to represent the functionalised aniline) (Figure 2-7). The reactants were initially used in a 1:1 stoichiometry with 5 equivalents of the $K_3Fe(CN)_6$ oxidising agent, to a final reactant concentration of 3 mM. After 3 h at room temperature, two new peaks were identified by HPLC in a 34:66 ratio (by HPLC peak area at 210 nm), alongside unreacted *p*-toluidine. Curiously, no unreacted 2-amino-*p*-cresol could be seen, suggesting that the *o*-aminophenol starting material was being consumed without cross coupling to the aniline. The new peaks were isolated and analysed by MS and the major species was identified as the desired product **24**. The side-product peak showed a + 16 Da increase in mass compared with the desired product, corresponding to an additional oxygen atom. It is proposed that this by-product results from dimerisation of the *o*-aminophenol starting material to give the iminoquinone **25**. With incomplete oxidation of the *o*-aminophenol species, this dimerisation could result from nucleophilic attack from the amine of one non-oxidised equivalent of *o*-aminophenol on one oxidised equivalent, in a similar manner to that proposed in the previous mechanism.

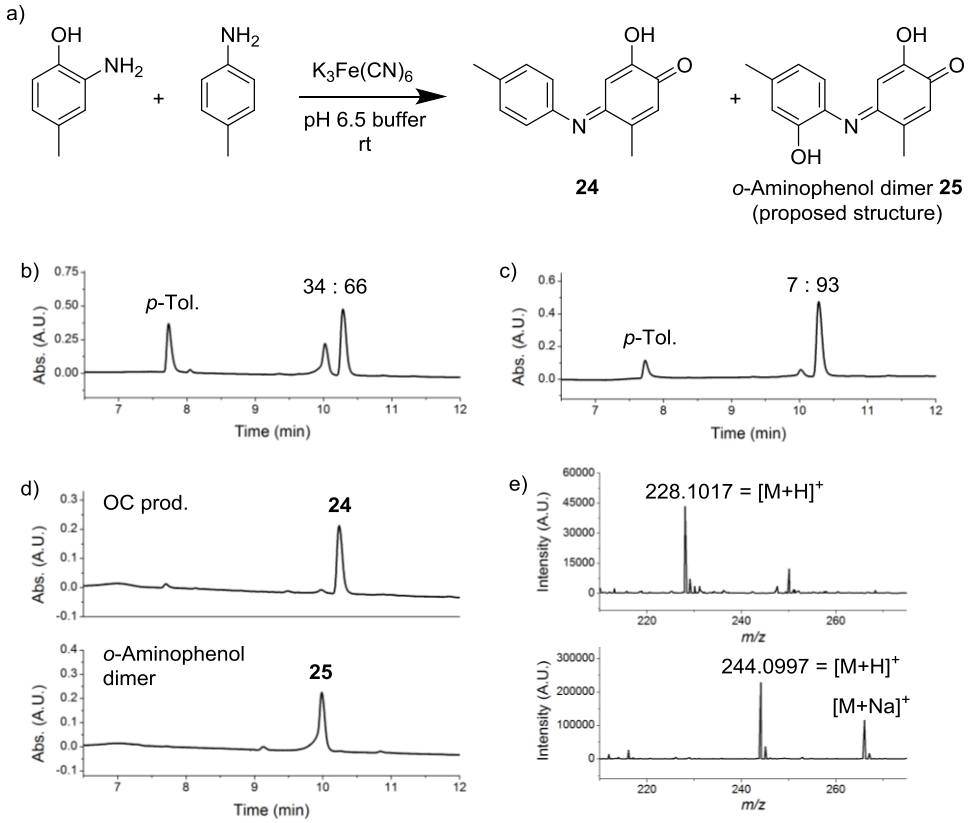


Figure 2-7 *Oxidative coupling of an o-aminophenol substrate. (a) The reaction conditions used with reactants in a 1:1 stoichiometry. (b) HPLC trace at 210 nm of the reaction mixture, showing a 34:66 ratio of dimer to product. Unreacted p-toluidine was also observed, but no unreacted 2-amino-p-cresol was seen (expected at 7.28 min). (c) Following adaptation of the conditions, the dimer to product ratio was reduced to 7:93. (d) Each of the peaks formed during the reaction were isolated. (e) MS of the isolated samples confirmed that the major peak corresponded to the desired product, while the minor peak appeared to result from dimerisation of the o-aminophenol reactant.*

The dimerisation of the *o*-aminophenol species was not expected to affect the final protein functionalisation step of the catch-and-release strategy, as protein modification is generally performed on low concentration samples with an excess of modifying reactant. Furthermore, the modified Tyr residues in a protein will be less reactive to dimerisation due to their comparatively lower solvent accessibility and fixed position in the protein sequence. However, conditions were explored to minimise dimerisation in the small molecule reaction. By doubling the number of equivalents of oxidant to 10, and performing the reaction at one third of the previous concentration (1 mM with respect to starting materials, previously 3 mM), not only did the reaction reach completion within 30 min but the ratio of dimer to product was reduced to 7:93 (by HPLC peak area at 210 nm).

The oxidative coupling strategy was next attempted using a fluorescent aniline in the form of the commercially-available fluoresceinamine and the Tyr-derivative **5** (Figure 2-8). Using the optimised conditions from the previous reaction, significant dimerisation of the Tyr-derived starting material was observed by HPLC, which may suggest that fluoresceinamine and/or Ac-Tyr(3-NH_2)-OMe **5** are less reactive in the oxidative coupling reaction than the reactants in the previous example. By increasing the number of equivalents of the aniline starting material, the ratio of dimer to product was significantly reduced (Table 2-2). These final conditions gave complete conversion by HPLC of the Tyr-derived starting material within 2 h. The product was identified by HPLC-PDA with a new absorbance at 442 nm, which was similar to that reported for the fluorescein moiety in acidic pH (at pH 4, λ_{ex} 437 nm)[125], at neutral pH the fluorescein moiety has been reported to have λ_{ex} 490 nm, λ_{em} 523 nm.[125]

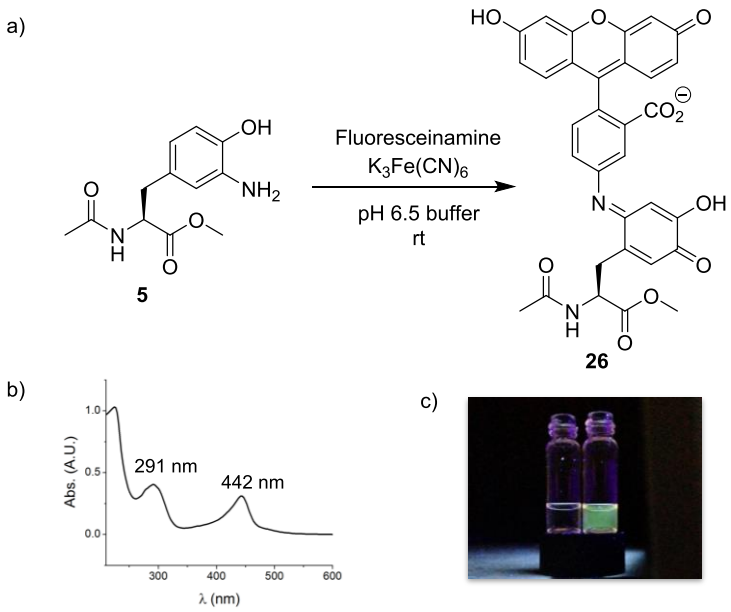


Figure 2-8 *Oxidative coupling of fluoresceinamine with the o-aminophenol modified Tyr residue. (a) The reaction conditions used to achieve oxidative coupling of fluoresceinamine to* ***5****. (b) PDA spectrum of the product shows incorporation of the fluorophore as a new absorbance at 442.2 nm appears. (c) Under irradiation with 365 nm light, the purified product (right) shows fluorescence.*

[Ac-Tyr(3-NH_2)-OMe] (mM)	[Fluoresceinamine] (mM)	[$K_3Fe(CN)_6$] (mM)	Dimer/Product ratio (HPLC at 210 nm)
1	1	10	45:55
1	5	10	16:84
1	10	10	11:89

Table 2-2 *Optimisation of the oxidative coupling reaction with fluoresceinamine as the nucleophile. The best ratio of dimer to product was achieved with 10 equivalents of fluoresceinamine compared to the o-aminophenol starting material. All reactions were performed at room temperature for 2 h, followed by desalting using a prepacked C_{18} SPE cartridge. Analysis was performed by HPLC-PDA at 210 nm absorbance.*

The oxidative coupling strategy has been shown to produce a fluorescently-labelled Tyr derivative, with scope for future optimisation for protein substrates through manipulation of the reactant and oxidant stoichiometry, and the final concentration with respect to the tyrosine-derived substrate. Crucially, this transformation was performed in near-neutral buffer at rt, avoiding the harsh and potentially denaturing conditions used in the diazotisation, azide formation, CuAAC approach. For this reason, this strategy was favoured for the post-cleavage functionalisation step of the catch-and-release modification strategy.

2.6 Conclusion

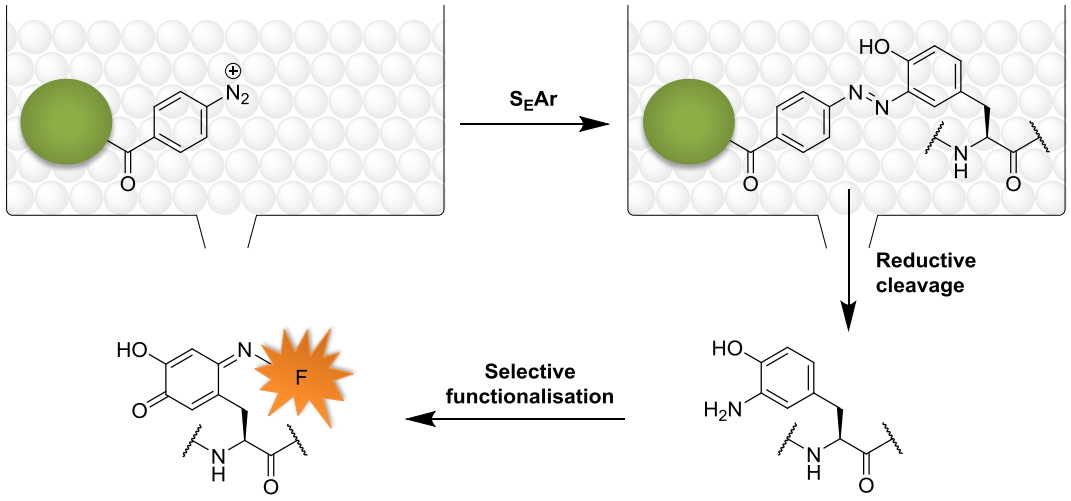
In this chapter, the three-step modification strategy for Tyr-bearing biomolecules was introduced. Although a solid-phase platform will ultimately be used to achieve selectivity among multiple Tyr residues, the chemistry of each step was first explored in-solution. To aid the development of this system, single amino acids derived from Tyr were synthesised, alongside a Tyr-containing peptide, Leu-enkephalin, which will be employed in the following chapter. In the development of the 'catch' step of the modification strategy, electrophilic aromatic substitution of phenolic substrates with a diazonium salt was demonstrated, and the regioselectivity of this transformation was confirmed by NMR. To 'release' the Tyr-bearing substrate, cleavage of the azobenzene linker is required, and this was assessed using aqueous sodium dithionite at various concentrations. Although complete reductive cleavage was observed within 15 min with even the lowest concentration trialled, it is expected that dithionite cleavage on-resin will require higher concentrations, as has been demonstrated in the literature.[66] Next, four approaches were considered for the post-cleavage functionalisation of the modified biomolecule. Amide bond formation with the aromatic amine was discarded due to a lack of selectivity, however, functionalisation with a click handle in the form of an azide was demonstrated *via* a diazotisation-azide formation approach. CuAAC was used to

furnish a Tyr residue bearing a fluorescent label. However this approach was ultimately considered to involve too many synthetic steps, while exposing the modified substrate to potentially denaturing conditions. As such, this strategy would be unsuitable for protein substrates, but may remain applicable to smaller peptides. Oxidation of the 3-aminotyrosine ring was next considered as a route to selective functionalisation. A Hetero-Diels-Alder reaction with an acrylamide reactant was ruled out due to the reported production of poorly defined bioconjugates,[120] however this approach led to the exploration of the oxidative coupling reaction which involved Michael addition of an aniline to the *o*-iminoquinone produced through ferricyanide oxidation of an *o*-aminophenol. The conditions of this reaction were optimised for two different anilines, which resulted in the production of a Tyr-derivative bearing a fluorescein label. This oxidative coupling strategy was considered the most promising for use with a protein substrate and so it was adopted for future work. After the establishment of each constituent reaction in the three-step Tyr-modification strategy, the next challenge was to move this onto a solid-phase platform, and this stage of the project will be explored in Chapter 3.

Chapter 3 Results & Discussion 2: Development of a solid-phase platform

3.1 The rationale for a catch-and-release approach

Using the chemistry discussed in Chapter 2, it should be possible to selectively label a protein substrate through the reactive Tyr residues. In order to achieve further selectivity for only the most exposed Tyr residues, the system was reproduced on a solid-phase where it was envisioned that S_EAr modification would occur *via* a surface-surface interaction between the protein and the resin-bound diazonium salt (Scheme 3-1).



Scheme 3-1 *An overview of the catch-and-release modification strategy using a solid-phase platform.* ***S_EAr:*** *The Tyr-bearing molecule is 'caught' on the solid-phase by S_EAr with an immobilised diazonium salt.* ***Reductive cleavage:*** *The modified biomolecule is released by reductive cleavage of the azobenzene linker, and bears a unique aromatic amine on the modified Tyr side chain.* ***Selective functionalisation:*** *Using an oxidative coupling strategy, selective conjugation of a fluorophore (F) is achieved.*

With this adaptation, the 'catch' and 'release' steps refer to immobilisation of the biomolecule on a solid-phase, and its subsequent release upon reductive cleavage of the linker. In addition to providing selectivity for exposed Tyr residues, a singly modified protein will become tethered to a bulky resin bead, limiting the opportunity for secondary modification. The use of a solid-phase platform serves a further role in allowing step-wise separation of diazotisation (which requires harsh acidic conditions) and introduction of the proteins substrate (in neutral pH buffer). Finally, the capture of a biomolecule on a solid-phase resin allows the straightforward separation of modified and unmodified substrate, as the unmodified material is washed from the resin prior to release of the modified substrate under reducing conditions.

3.2 Streptavidin resin

The first solid-phase platform that was considered was resin-bound streptavidin (SAv) which, once functionalised with an aniline for diazotisation, would provide significant steric bulk around the diazonium salt in order to favour exposure-based selectivity. SAv resin is commercially available and consists of cross-linked agarose beads functionalised with the ~56 kDa SAv protein *via* a covalent linker (Figure 3-1).[126, 127] The exact nature of this linker is not explicitly described by the vendor, but it is most likely constructed through formation of an amide bond with the side chain of a Lys residue in the protein. SAv exists as a tetramer, with each subunit binding one biotin molecule, and as there is some cooperativity between SAv subunits during biotin binding, it is assumed that the protein is immobilised in its tetrameric form.[126]

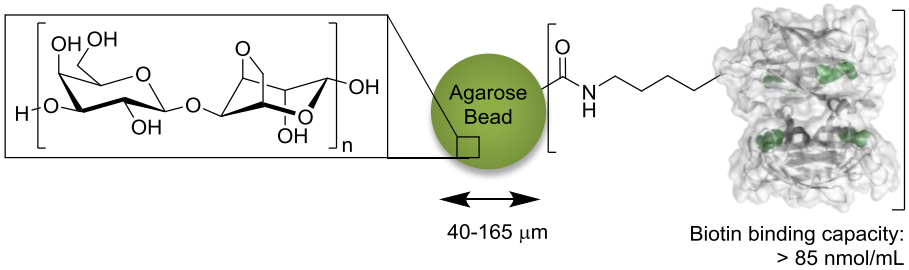


Figure 3-1 *SAv-functionalised agarose resin was used as the solid-phase platform.* ***(Left)*** *The resin is composed of 6% crosslinked agarose, which contains the D-galactose and 3,6-anhydro-L-galactopyranose subunits.* ***(Right)*** *A streptavidin tetramer is shown with biotin ligands bound to it (green spheres). The protein is most likely immobilised on resin via a covalent linkage with a Lys residue, forming an amide linker (PDB ID: 1SWE).*

SAv is structurally similar to Avidin (Av), and both of these proteins bind biotin with high affinity (SAv: $K_d \approx 10^{-14}$ M; Av: $K_d \approx 10^{-15}$ M).[127] Although SAv binds biotin with slightly lower affinity than Av, it was favoured for the current application as, in contrast to Av, it does not bear any carbohydrate post-translational modifications (PTMs) which are thought to cause non-specific protein-protein binding. In order to use SAv resin in the catch-and-release system, it first required functionalisation with an aniline from which a diazonium salt could be generated. SAv functionalisation was achieved through derivatisation of the biotin ligand, which is a strategy that is commonly employed in Chemical Biology for 'pull-down' experiments of biotinylated substrates from complex mixtures.[128] As such, through the synthesis of a biotinylated aniline, functionalisation of the resin could be achieved by incubation of the modified ligand with the SAv resin.

3.2.1 Synthesis of a biotinylated aniline

When biotin is bound to the SAv binding site, the ureido headgroup of the biotin ligand is largely buried and is involved in cooperative H-bonding with five residues in the protein sequence (Figure 3-2).[129] However, the valeric acid side chain points out of the binding pocket towards the solvent and is therefore relatively accessible. This site therefore represents an attractive target for derivatisation using straightforward amide bond forming chemistry.

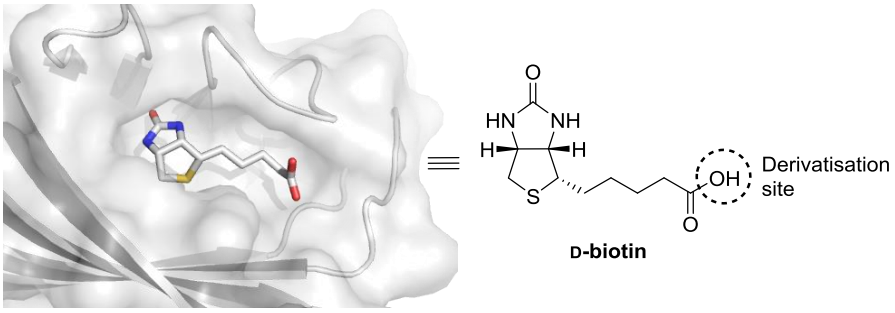
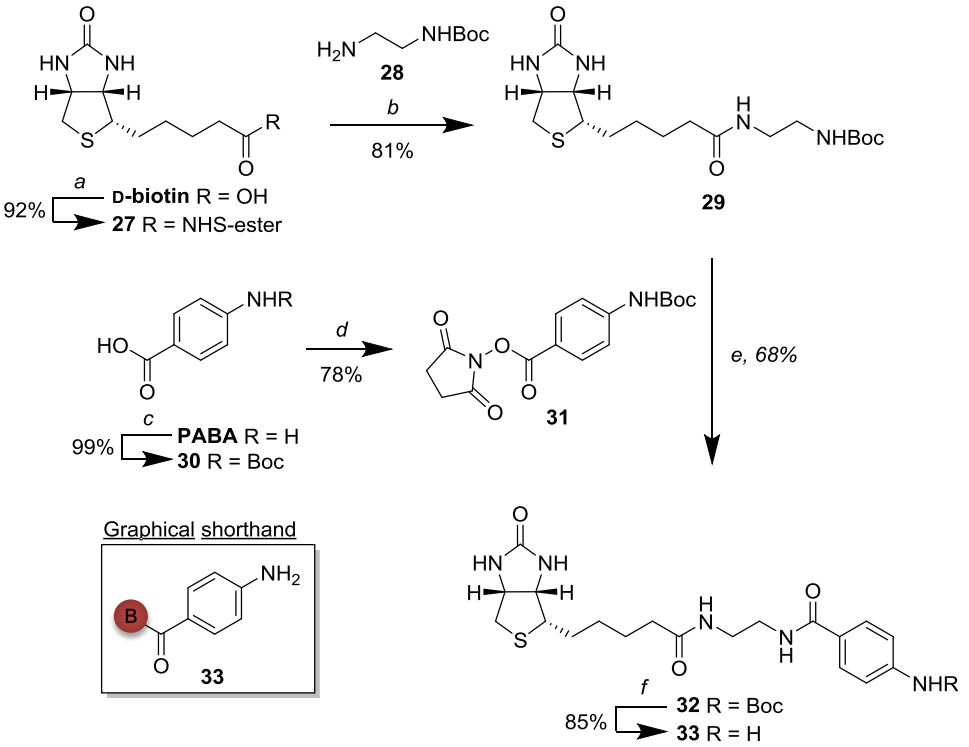


Figure 3-2 *The organisation of biotin in the binding site of SAv. The urea group is the most deeply buried region of the biotin ligand, while the carboxylic acid is the most solvent-exposed (PDB ID: 1SWE).*

The biotinylated aniline **33** was synthesised starting from D-biotin, and the terminal carboxylic acid was first activated as an NHS-ester using *N*-hydroxysuccinimide and DCC in DMF at room temperature. Due to the poor solubility of the NHS-activated biotin derivative **27**, the product was purified by trituration with ether, after removal of DMF *in vacuo*. This purification method avoids the use of flash chromatography, which may cause hydrolysis of the NHS-ester. A linker was next installed between the biotin and aniline moieties using a short Boc-protected diamine **28**, which resulted in formation of the biotin derivative **29**.



Scheme 3-2 *Synthesis of a biotinylated Boc-protected aniline for application in the catch-and-release system. (a) NHS, DCC, DMAP, DMF, rt, 18 h; (b)* ***28****, NEt_3, DMF, rt, 18 h; (c) Boc_2O, NEt_3, Dioxane/H_2O (2:1), rt, 18 h; (d) NHS, DCC, DCM, rt, 18 h; (e)i) TFA, DCM, rt, 2 h; ii) $NaHCO_3$,* ***31****, MeCN, rt, 18 h; (f) TFA, DCM, rt, 2 h.*

A Boc-protected aniline with an NHS-ester in the *para* position was next synthesised in two steps from *para*-aminobenzoic acid (PABA). The aniline was first protected using Boc_2O to give **30**, before activation of the carboxylic acid as an NHS-ester. The resulting amine-reactive reagent **31** was once again purified without the use of flash chromatography to avoid hydrolysis of the NHS-ester. Deprotection of the biotin derivative **29** was achieved in TFA with monitoring by TLC with ninhydrin staining. After removal of the solvent and TFA under a stream of nitrogen, the residue was redissolved in saturated sodium bicarbonate, before introduction of **31**. The resulting Boc-protected biotinylated aniline **32** was then deprotected in TFA/DCM to give the desired biotinylated aniline **33** (4 steps from D-biotin, 43% overall yield). With the necessary reagent in hand to equip SAv resin with an aniline, diazotisation and azobenzene formation was attempted, first in-solution before moving onto the solid-phase.

3.2.2 Azobenzene formation and unforeseen biotin modification

To prepare the biotinylated aniline **33** for azobenzene formation with a Tyr-derived substrate, the aniline was first diazotised in acidic conditions with $NaNO_2$, before neutralisation and introduction of Ac-Tyr-OMe **2** to form the S_EAr product **13** (Figure 3-3). Suspicions were aroused that this product was not the intended biotinylated azobenzene, due to analysis by ESI-TOF MS, which did not agree with the expected mass of the product. In fact, instead of the expected $[M+H]^+$ signal at 654.2704, *m/z* signals were observed at 653.2603, 683.2605 and 705.2426 (in a sample which gave a single peak by HPLC). The signals at 683.2605 and 705.2426 were particularly interesting as they appeared to correspond to a single species bearing a proton and sodium adduct respectively, the 653.2603 signal presumably resulted from degradation of this species.

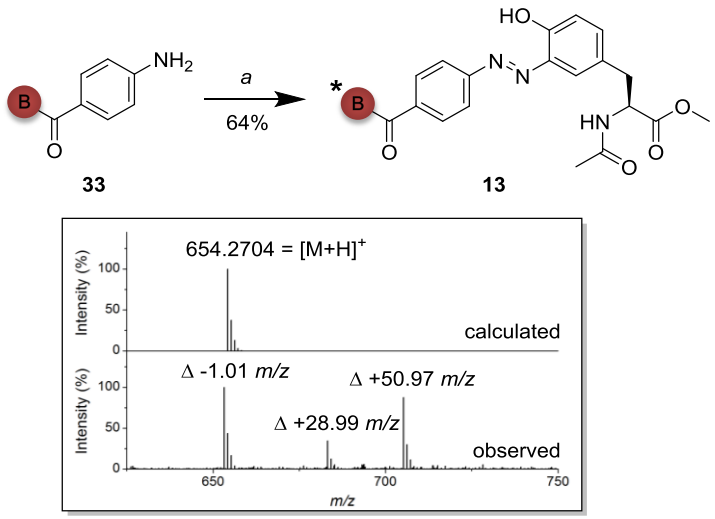


Figure 3-3 *Azobenzene formation with a biotinylated diazonium salt gave rise to an unknown product* ***13****. (a)i) HCl, $NaNO_2$, H_2O, 0 °C, 20 min; ii)* ***2****, $NaHCO_3$, 0 °C – rt, 15 min. Inset: the unknown masses observed by MS, compared to the theoretical mass.*

In order to explore this further, NMR studies were performed in DMSO-d_6 due to the poor solubility of the biotinylated product **13**. Using an NMR solvent which does not cause H/D exchange of labile protons became important, as it emerged that only one N*H* from the biotin urea group could be seen in the azobenzene product **13**, shifted downfield from those in D-biotin from ~6.3–6.4 ppm to 8.80 ppm. The MS and NMR results were best explained by addition of a nitrosyl group to the biotin urea, resulting in the unintended nitrosylated product **13** (Figure 3-4).

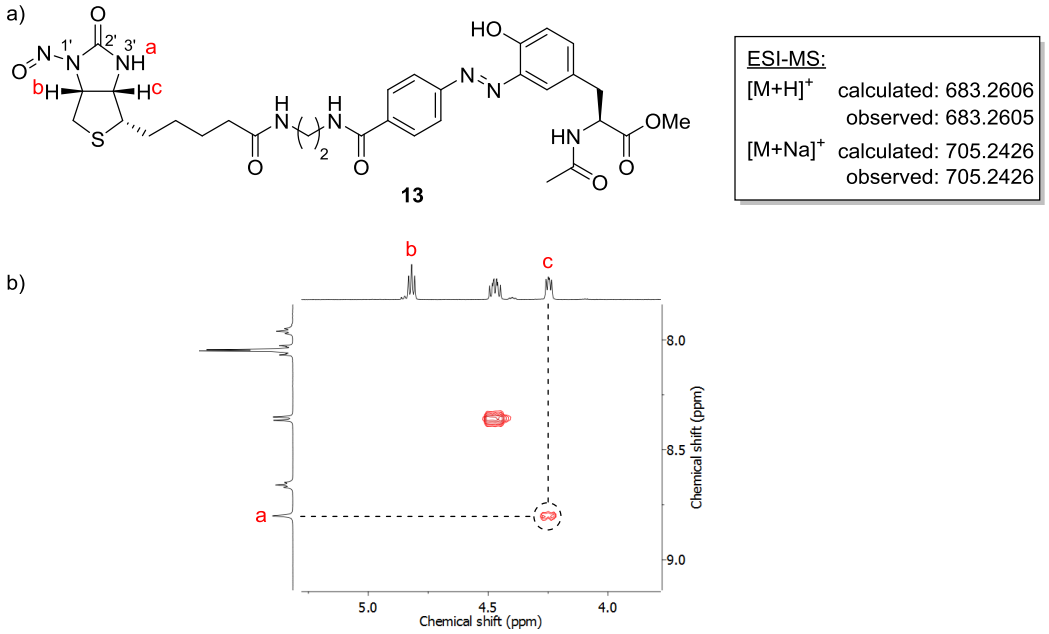


Figure 3-4 *Nitrosylation of the biotin moiety under azobenzene forming conditions. (a) The addition of NO was supported by HRMS. (b) By COSY NMR, only one urea NH was identified, and this was found to be located on the same side of the fused bicycle as the valerate chain [N(3')H], suggesting modification at the N(1') position.*

To determine the regiochemistry of the nitroso- modification, the compound was characterised by COSY NMR. There was a strong COSY interaction between the remaining urea-N*H* and the biotin C*H* adjacent to the valerate chain, suggesting that the modified site must be on the opposite side of the fused ring system [N(1')]. This finding was further explored using D-biotin as a nitrosylation substrate.

3.2.3 Biotin nitrosylation

Biotin nitrosylation is not a well reported transformation despite the importance of biotin in biological systems and the common use of both biotin and azobenzene chemistry in Chemical Biology. However, nitrosobiotin is not entirely unknown, as two papers (Jansen and Stokes in 1962 and Sanzini *et al.* in 1977) were found which describe its preparation using $NaNO_2$ in acidic conditions.[130, 131] Unfortunately, the characterisation of nitrosobiotin was limited to UV/Vis, IR spectroscopy and elemental analysis, and so there was no determination of the regiochemistry of the transformation. Beyond this, a passing mention of a '1-nitrosoimidazole product' was reported as an unintended by-product by Welzel *et al.* when they attempted to modify an antibiotic natural product with a biotin ligand using diazonium chemistry.[131] Although this statement appeared to assign the modification to the N(1') position, the only supporting data provided was ^{13}C NMR and MS, which did not allow regiochemistry to be established. The justification for modification at the N(1') position may be through analogy to

carboxybiotin, as it is the N(1') position which is modified in biotin during its role in CO_2 transfer.[133] However, as there is no published data to support the regiochemistry of biotin nitrosylation, it was explored using modern analytical techniques.

D-Biotin was successfully converted to nitrosobiotin **34** through *in situ* generation of the nitrosonium ion using TFA and $NaNO_2$. The reaction was performed at room temperature and was judged complete by UHPLC after 2 h (Figure 3-5). As before with the modified biotin substrate, the remaining N*H* of the biotin urea was identified at 8.80 ppm, and was found by COSY NMR to correspond to the N(3') position.

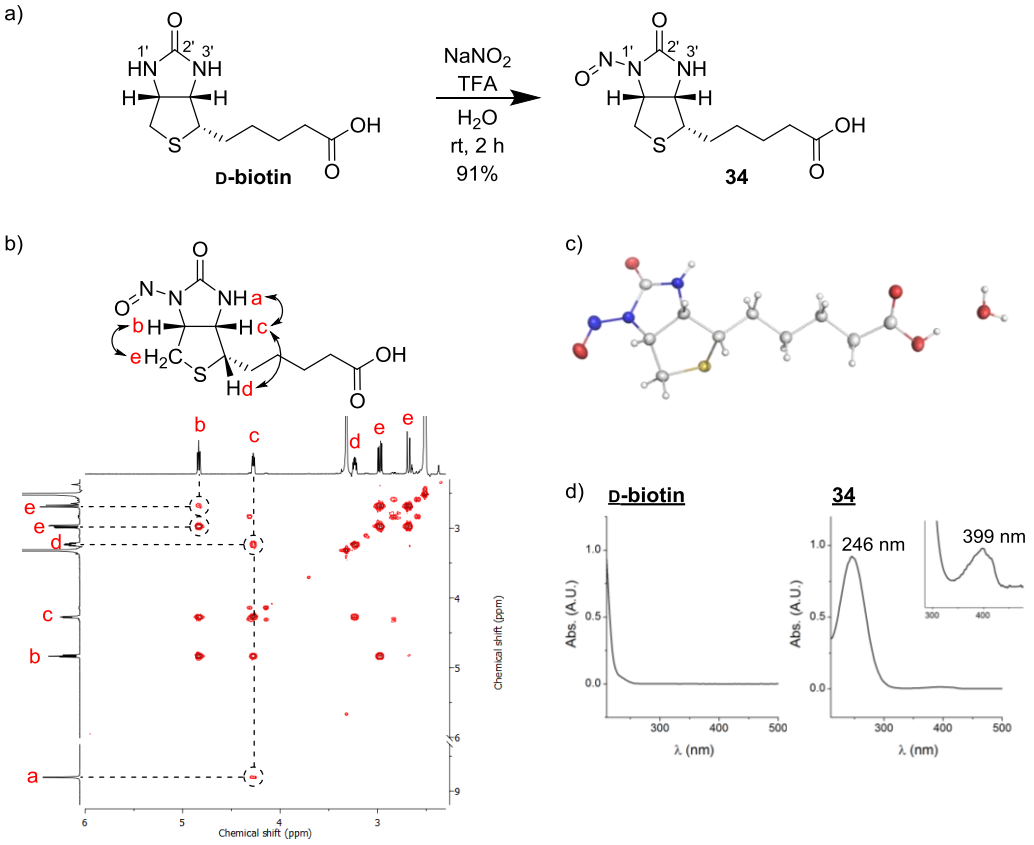


Figure 3-5 *The nitrosylation of* D*-biotin. (a) The reaction conditions which give rise to nitrosobiotin* ***34****, the key atoms of the biotin urea group are labelled by the convention used by Strzelczyk et al.*[134] *(b) By COSY NMR, the nitrosylated site was found to be N(1'), as the remaining NH was identified on the same side as the valeric acid chain. (c) The X-ray crystal structure of nitrosobiotin confirmed the location of modification. (d) By HPLC-PDA, the nitroso-modification was found to cause two new absorbance maxima at 246 and 399 nm.*

An X-ray crystal structure was obtained of nitrosobiotin **34**, following recrystallisation from H_2O:EtOH (1:1). This provided further evidence that biotin was modified at the N(1') position, although it was not immediately clear why there would be a difference in reactivity between the N(1') and N(3') sites. There is no significant difference in the electronics of these positions,

and despite the presence of the valeric acid chain on the N(3') side, this is shown in X-ray structures as an elongated chain which would not seem to provide sufficient steric interference to reduce the reactivity of the N(3') position. However, work by Mazurek *et al.* has found through *in silico* calculations that a folded conformation of biotin (in which the valeric acid chain bends back towards the urea group) may be stabilised by intramolecular H-bonding between the N(3') position and the terminal carboxylic acid (Figure 3-6).[134]

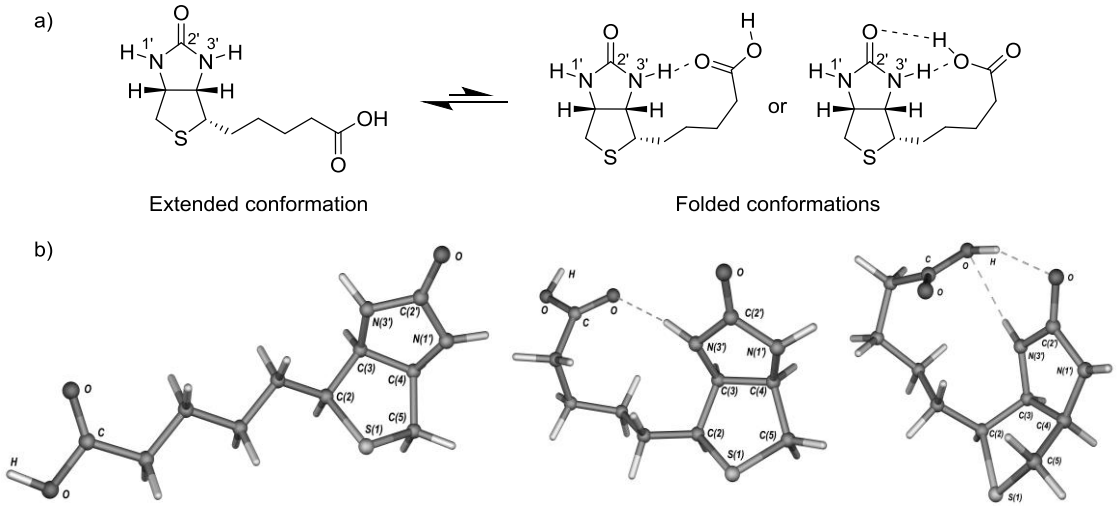
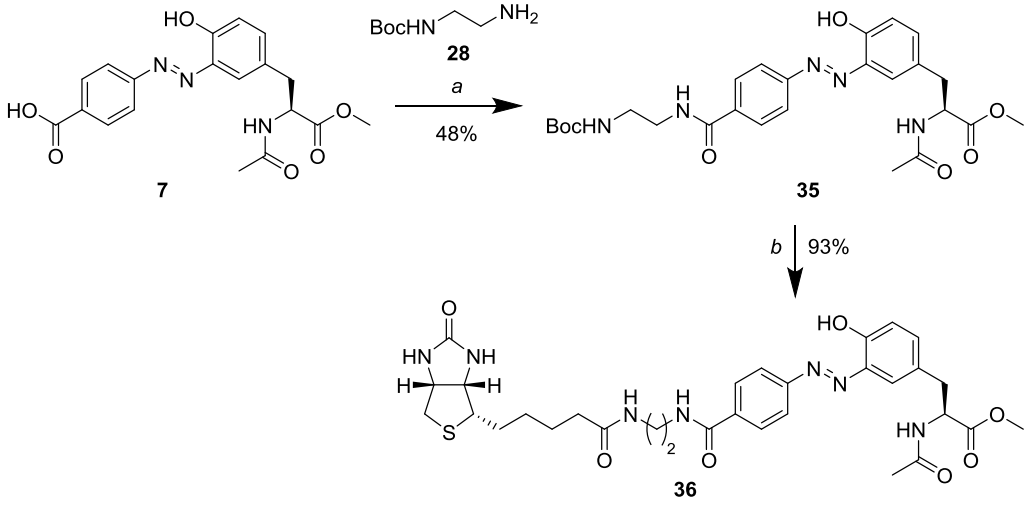


Figure 3-6 *The conformations of biotin. (a) In X-ray crystal structures, the extended form of the valeric acid chain is stabilised by intermolecular H-bonding. However when biotin was modelled in a vacuum, intramolecular H-bonding between the N(3') position and the terminal carboxylic was found to stabilise folded conformations. (b) The optimised structures reported by Mazurek et al. Images reproduced with permission from the publisher.*[134]

Further evidence to support the existence of the proposed folded biotin conformations was described by Li *et al.*[135] Using quantum chemical calculations to assess the different rates of H/D exchange at the N(1') and N(3') positions, the researchers found that the N(1') position had an exchange rate 6 times greater than that of the N(3') position, and this was supported by experimental data. This result was further explained by the finding that the relative proportion of extended, semi-folded and folded biotin in water was 24:3.6:1.[135] While the proportion of folded biotin was found to be low, the folding and unfolding of the valeric acid chain is a dynamic process, and therefore the different reactivities of the N(1') and N(3') positions can be ascribed to the fact that all biotin molecules in the sample will spend some time in the folded conformation. This work serves as a useful reminder that the X-ray structure of a crystalline material represents its conformation once fixed in space, and may not capture the dynamic conformational changes which occur in the solution-phase. With the rise of computational chemistry, modelling these solution-phase changes should become more commonplace, and

hopefully this will provide data on molecular structure which will complement that found by X-ray crystallography.

In order to obtain a sample of the intended biotinylated linker without the nitroso-modification, the target compound **36** was synthesised by performing azobenzene formation prior to attachment of the biotin label (Scheme 3-3).



Scheme 3-3 *Synthesis of biotinylated Tyr **36** which avoids biotin nitrosylation. (a) **28**, EDC, HOBt, DMF, rt, 18 h; (b)i) TFA, DCM, rt, 3 h; ii) Biotin-NHS **27**, sat. $NaHCO_3$, DMF, rt, 18 h.*

The synthesis of **36** essentially reverses the order of steps used for the synthesis of nitroso-modified **13**, starting from the PABA-derived azobenzene **7**. Using amide coupling reagents, this compound was modified with the Boc-protected diamine **28**, before deprotection and amide coupling with biotin-NHS **27**. The product was confirmed by NMR and MS as **36**, the biotinylated, azo-linked Tyr derivative which did not carry the nitroso modification.

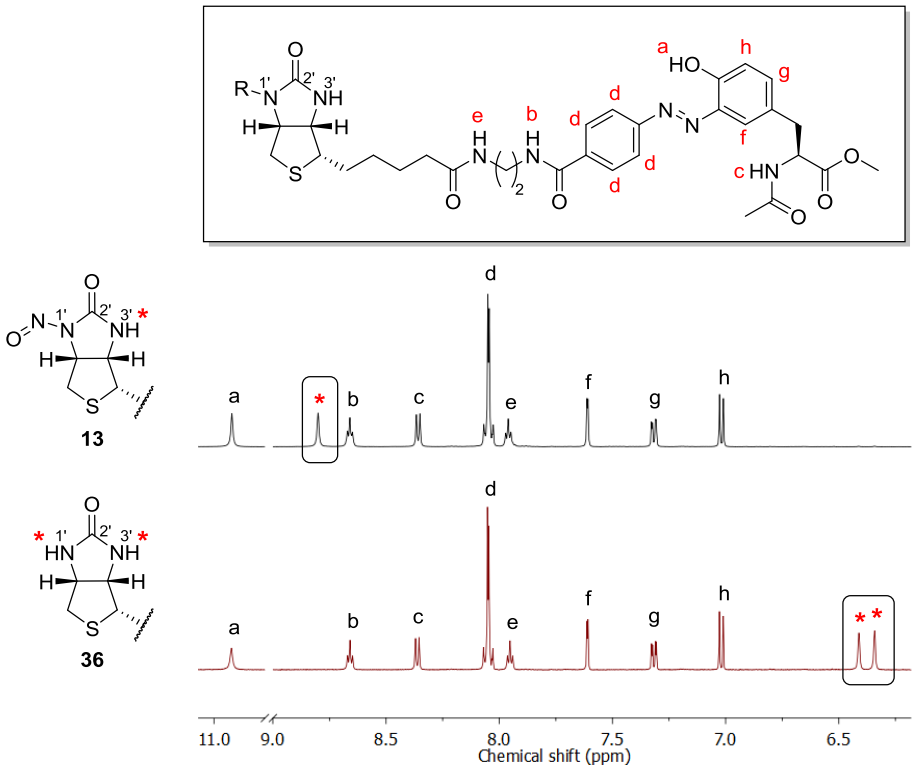


Figure 3-7 *Comparison of the ^{1}H NMR of nitrosylated biotin linker **13** and the biotin linker **36**. In the nitrosyl-modified molecule, the N(3')H of the nitrosourea group (*) exhibits a diagnostic chemical shift at 8.80 ppm. The biotin linker displays two peaks at 6.41 and 6.34 ppm, corresponding to the urea NH protons (*).*

By ^{1}H NMR in DMSO-d_6, the distinction between **13** and **36** was clear when considering the urea N*H* peaks (Figure 3-7). As described previously, the N(3')*H* peak in nitroso-containing **13** was observed at 8.80 ppm as a singlet which integrated to 1 proton. By comparison, the two urea N*H* peaks in **36** were identified at 6.41 and 6.34 ppm, each presenting as a singlet which integrated to 1 proton. The downfield shift of the remaining urea peak following nitrosyl modification confirms its electron-withdrawing properties, which cause deshielding of the N(3') position.

In an unexpected result, it was found that the nitrosylated biotin linker **13** retained the ability to bind to SAv resin, allowing the immobilisation of an azo-linked Tyr derivative onto a solid-phase for the first time in this project. Sub-stoichiometric portions of the preformed linker (3 nmol in 1 mL) were introduced to 0.2 mL of SAv resin and incubated at room temperature for 15 min with agitation on a rotator. The eluate was then drained from the resin and analysed by UHPLC-PDA with observation at 330 nm, corresponding to the absorbance of the azobenzene group [Figure 3-8 (a)]. The expected binding capacity of the resin was >17 nmol of biotinylated material,[1] and it was found that after 6 × 3 nmol portions had been incubated with

the resin, the amount of unbound ligand in the eluate hit a maximum and began to plateau, indicating that there was no further binding of **13** to the immobilised protein.

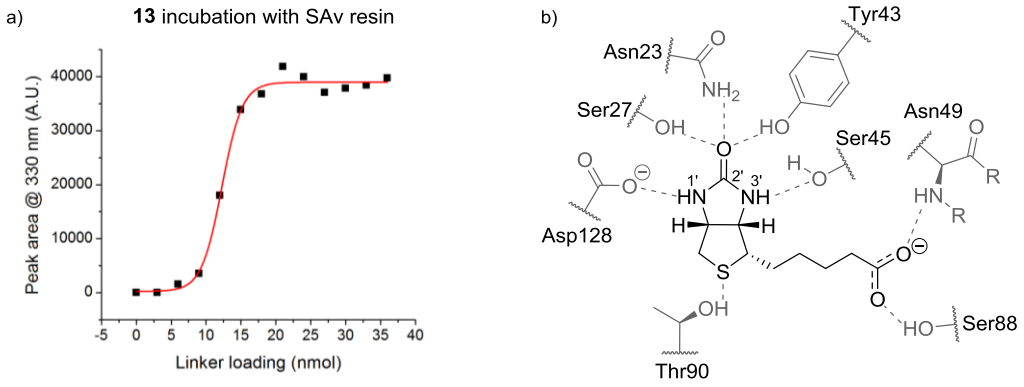
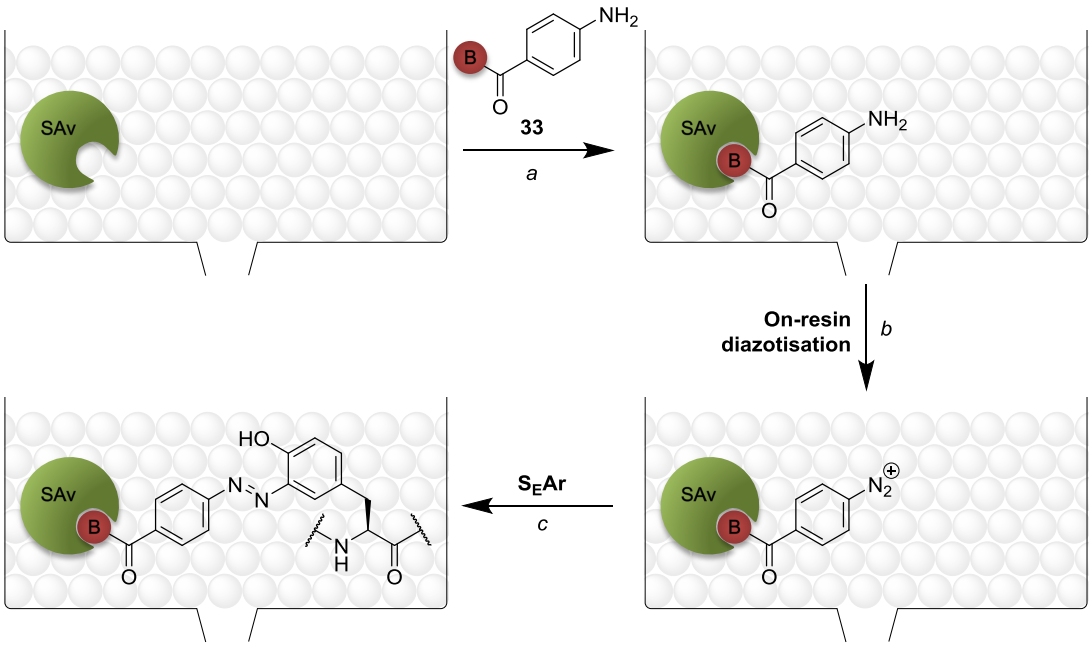


Figure 3-8 *Binding of biotin derivatives to SAv. (a) Despite modification of the urea group, the nitrosylated biotin linker* ***13*** *showed the expected binding curve when incubated with SAv resin. 3 nmol aliquots were incubated with 0.2 mL resin for 15 min at rt. Unbound ligand was washed from the resin and analysed by UHPLC. After ~18 nmol, the SAv binding sites appeared saturated. Expected resin capacity was >17 nmol for 0.2 mL resin (> 85 nmol/mL). (b) The interactions between biotin and key residues in the SAv binding pocket are shown.*[136]

While the nitrosylated linker appeared to bind to SAv, it is unknown how the binding affinity of nitrosylated derivatives would compare with the native ligand. By examination of the biotin binding site in SAv [Figure 3-8 (b)], there are several H-bonding interactions which give rise to the high binding affinity of the biotin ligand. In particular, the urea group appears crucial to stabilisation of the bound ligand, with five interactions highlighted between SAv residues and the urea moiety. Nitrosyl modification of the N(1') position will likely affect these interactions through increased steric bulk and alteration of the electronics of the urea group [as evidenced by the downfield shift of the N(3')*H* NMR signal described previously]. The final iteration of the catch-and-release method will involve diazotisation of the aniline after immobilisation on the SAv resin, so it is not clear whether the nitrosyl modification will still occur when the biotin headgroup is already bound to the protein. However, given the apparent ability of both the nitrosyl-modified and unmodified biotin substrate to bind to SAv, the catch-and-release method was attempted on SAv resin, initially using small molecule and peptide substrates.

3.2.4 Catch-and-release using SAv resin

In order to activate the SAv resin for Tyr capture, it was functionalised with an aniline using the biotinylated reagent **33** [Scheme 3-4 (a)]. Following aniline functionalisation, on-resin diazotisation was required to activate the resin for Tyr-specific capture [Scheme 3-4 (b)]. Finally, with introduction of the Tyr-bearing substrate, azobenzene formation would occur upon S_EAr modification of the Tyr phenol ring [Scheme 3-4 (c)].



Scheme 3-4 *Overview of the SAv resin-based modification strategy. (a) The biotinylated aniline **33** was loaded onto SAv resin in neutral pH buffer. (b) The immobilised aniline was next diazotised using $NaNO_2$ in acidic conditions. (c) Upon introduction of a Tyr-bearing substrate, S_EAr occurs, resulting in formation of an azobenzene group.*

Aniline functionalisation of the resin was performed by addition of biotinylated aniline **33** in buffer (pH 7.2, 10% DMSO for solubility) with agitation on a rotator for 2 h at room temperature. After aniline loading, the resin was washed with buffer until the eluate was found to contain no aniline by UHPLC. Initial attempts at on-resin diazotisation employed HCl or TFA with $NaNO_2$ to generate the reactive nitrosonium ion *in situ*, however it was observed that this resulted in release of the biotinylated aniline, presumably due to denaturation of the SAv protein, or decomposition of the resin itself.

In order to achieve diazotisation under milder conditions, acetic acid was used (0.6 M, ~pH 2.5). After removal of the diazotisation reagents by vacuum filtration, Ac-Tyr-OMe **2** was introduced to the resin in buffer (pH 7.2, 10% DMSO for solubility). The mixture was agitated on a rotator at room temperature for 18 h, during which time the resin developed an orange

colour due to the absorbance of the azobenzene linker. Excess Ac-Tyr-OMe was removed by washing the resin and the eluate was analysed by UHPLC until no starting material remained. An aqueous solution of sodium dithionite was then introduced and the resin was agitated at room temperature for 18 h. The eluate was analysed by UHPLC, and a new peak was identified which was distinct from the unmodified substrate [Figure 3-9 (a)]. The cleavage product was identified in a qualitative manner by matching retention time and PDA spectrum with a synthetic standard of Ac-Tyr(3-NH_2)-OMe **5**.

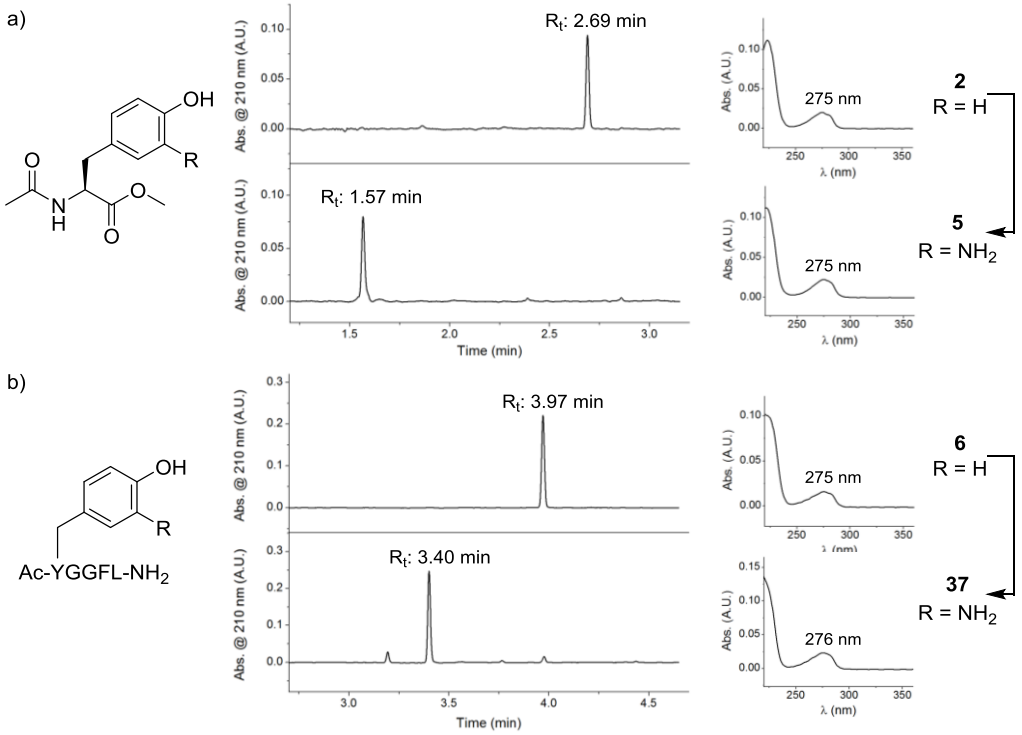


Figure 3-9 *Catch and release of small molecule and peptide substrates using SAv resin. Ac-Tyr-OMe* ***2*** *(a) and Ac-Leu-enkephalin-OMe* ***6*** *(b) were successfully modified using the activated SAv resin, the o-aminophenol modified product was identified by UHPLC.*

Using the Leu-enkephalin-derived substrate **6**, catch-and-release modification was performed as before, with identification of the cleavage product by UHPLC (Figure 3-9b), by comparison with a synthetic standard **37**. In both cases, the cleavage product had a shorter retention time on the reverse-phase UHPLC column than the unmodified substrate, due to the addition of an amine which increases the overall polarity of the catch-and-release product, however the PDA absorbance spectrum was largely unchanged by the modification.

3.2.5 Myoglobin as a protein substrate

Horse heart myoglobin (Mb) was identified as a suitable protein substrate for catch-and-release modification as it contains two Tyr residues in the native sequence. In addition, Mb is commercially available and inexpensive, and is frequently used as a calibrant for MS instruments, so it was known to ionise well in ESI-MS.

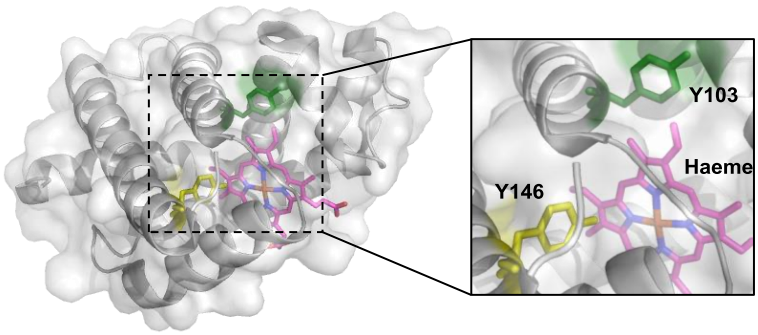


Figure 3-10 *The structure of horse heart myoglobin (PDB ID: 5D5R). There are two Tyr residues in the native sequence of the protein: Y103 (green) whose side chain is surface-exposed, and Y146 (yellow) which has the phenol group pointing towards the protein interior. Myoglobin also has a haeme group which is shown in pink.*

By retrieval of the Mb X-ray crystal structure from the Protein Data Bank (PDB), the two Tyr residues Y103 and Y146 were visualised, alongside the iron-containing haeme group which gives the protein its red colour (Figure 3-10). While both Tyr residues were found close to the protein surface, the reactive *ortho* positions of Y103 pointed outward towards solvent, while the Y146 side chain pointed towards the protein interior. By RP-HPLC, it was found that the haeme group becomes dissociated from the protein under the denaturing conditions of the LC system, and is observed as a separate peak with an absorbance at 398 nm (Figure 3-11).

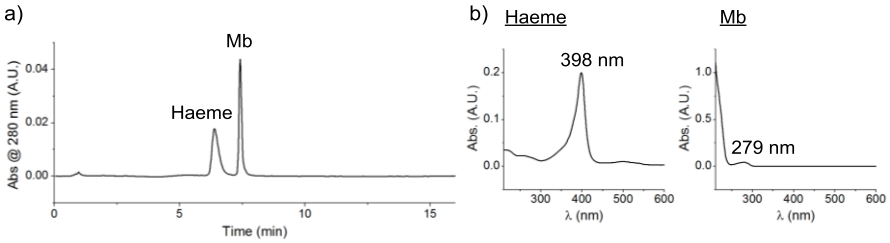


Figure 3-11 *RP-HPLC of horse heart myoglobin. (a) The haeme group was dissociated from the protein and these were observed as distinct peaks by HPLC (observation at 280 nm). (b) The PDA absorbance spectra of the haeme peak and the Mb protein peak.*

3.2.6 Catch-and-release of Mb

The protocol for the catch-and release reaction was adapted slightly for use with a protein substrate. To allow for analysis by MS, a size exclusion spin cartridge was used to desalt the cleavage product. Otherwise, the protocol remained largely unchanged, and UHPLC analysis with a C_4 column was used to follow the progress of the wash and cleave steps. Following reductive cleavage, the protein was analysed by FT-ICR MS. The released protein sample was found to contain primarily unmodified Mb, with additional peaks attributed to the unmodified protein bearing a phosphate adduct [Figure 3-12 (a)]. A signal corresponding to the expected modified protein was identified, but this appeared at <2% relative intensity when compared with the unmodified species. The haeme group was also identified in the mass spectrum, giving a mass which corresponded to $[M]^+$ [Figure 3-12 (b)]. After consulting the literature, this was identified as the haeme group bearing a positive charge with no additional adduct, and is known to arise from oxidation of the central Fe atom from the 2+ to the 3+ oxidation state.[137]

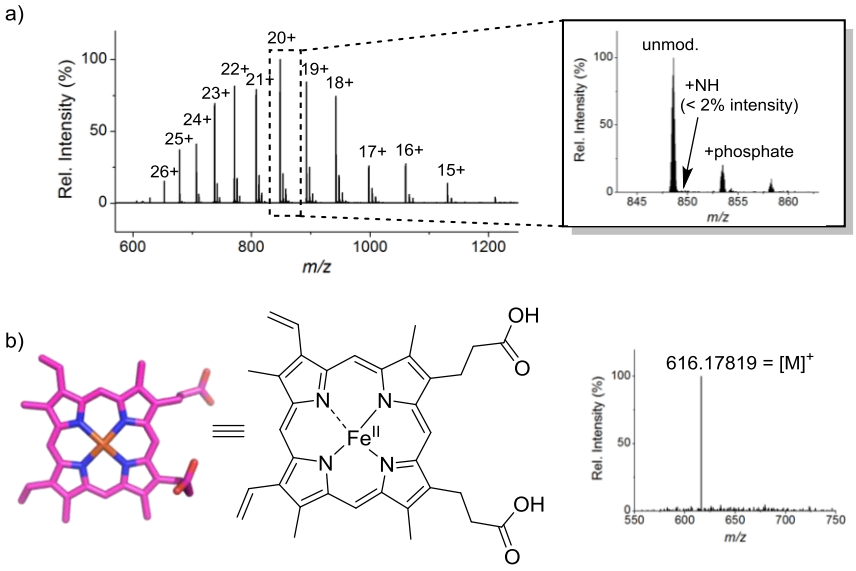


Figure 3-12 *Catch and release of a protein substrate. (a) By FT-ICR MS, the released protein sample was found to primarily consist of unmodified Mb. A small amount of modified protein was seen at less than 2% relative intensity. (b) The dissociated haeme group was identified as the* $[M]^+$ *ion, also known as ferri-haeme, which has been reported to arise from oxidation of the iron centre to* Fe^{3+}.

MS analysis of the Mb catch-and-release product had shown that there was very little modified protein present in the sample. This result was initially puzzling, as the eluate from the wash step had been monitored by UHPLC, and the resin was washed until clear of protein. Upon addition of dithionite, the protein peak had then reappeared. While SAv resin had been selected to provide steric bulk, it was concluded that there may have been non-specific binding between Mb and SAv during the experiment, giving rise to retention of Mb on the resin without the

expected modification of the substrate. Presumably, this non-covalent interaction was then disturbed by the increased salt concentration upon addition of dithionite, leading to release of Mb without the +15 Da change in mass expected from the amine modification. The reported isoelectric point of SAv is ~6, therefore it would have a slightly negative charge under the reaction conditions of pH 7.2.[138] Mb, by comparison, has an isoelectric point ~7 and is likely to have close to a neutral net charge under reaction conditions.[139] It is quite possible that the negatively charged SAv protein is responsible for the non-specific retention of the protein substrate, and while a nearly neutral Mb has shown retention on the resin, this effect is likely to be exacerbated in proteins with a higher isoelectric point, which would bear a net positive charge under reaction conditions. The proposed benefit of using a readily-available immobilised protein as the catch-and-release platform was ultimately undermined by the complications of non-specific binding and so, in order to avoid this effect with future protein substrates, a simplified resin was sought which did not bear an immobilised protein.

3.3 Affi-Gel 10 resin

Affi-Gel 10 resin by Bio-Rad was selected as a suitable, non-protein bearing solid-phase for the next iteration of the catch-and-release modification strategy. As with the SAv resin, the beads are composed of cross-linked agarose, which consists of two sugar subunits, D-galactose and 3,6-anhydro-L-galactopyranose (Figure 3-13). Rather than relying on functionalisation of the resin *via* the SAv-biotin interaction, an NHS-ester functionalised variant of the agarose gel was used, which allows modification of the resin through a straightforward amide bond-forming reaction. Documentation supplied by Bio-Rad suggests that the agarose resin is equipped with an NHS-ester through an ether linkage with a hydroxyl group on one of the sugar subunits.[140]

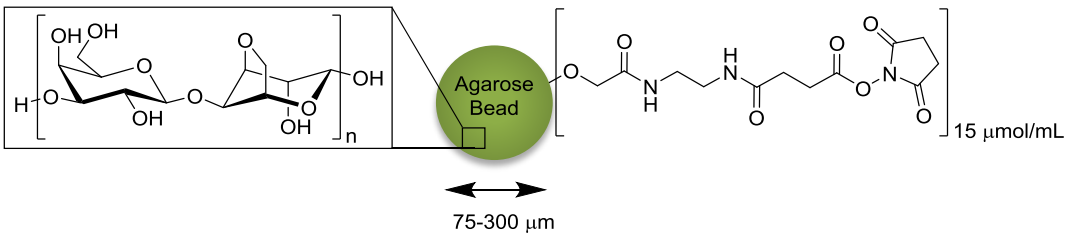
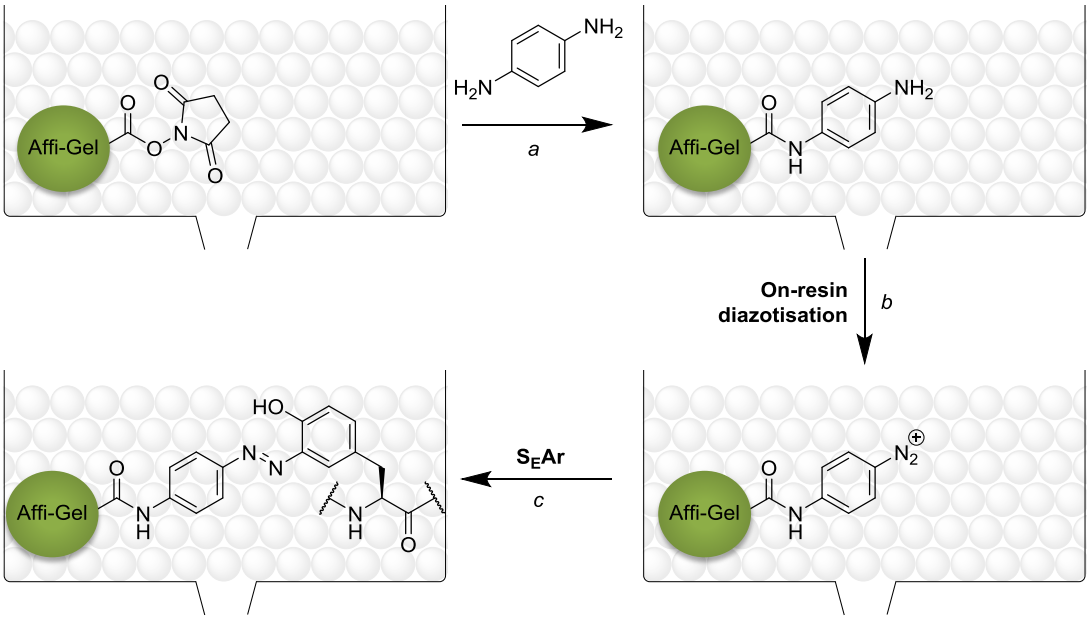


Figure 3-13 *Affi-Gel 10 resin from Bio-Rad. The reported average diameter of a polymerised agarose bead is shown, alongside the structure of the polymer subunits (D-galactose and 3,6-anhydro-L-galactopyranose) (**Left**) and the loading of the NHS-ester functional group (**Right**), which is attached via an ether linkage to a hydroxyl group on the cross-linked polymer.*

Benefits of the move to Affi-Gel resin include an increased substrate binding capacity of 15 µmol/mL (from >85 nmol/mL biotin binding with the SAv resin). In addition, the new resin could be equipped with an aniline in one step using a commercially available diamine, thus avoiding the synthesis of biotinylated linkers which were required in the previous work.

3.3.1 Functionalisation of Affi-Gel resin

In order to accommodate the new solid-phase, the catch-and-release protocol was adapted as shown in Scheme 3-5. The resin was equipped with the required aniline for diazotisation by incubation with an excess of *p*-phenylenediamine, the simplest symmetrical aromatic diamine. This reagent was commercially available and inexpensive, and was used in excess in order to avoid the need for an amine protecting group strategy.



Scheme 3-5 *Functionalisation of Affi-Gel 10 resin for use in the catch-and-release system. (a) The NHS-ester functionalised resin was equipped with an aniline using the symmetrical diamine p-phenylenediamine. (b) Diazonium salt formation using $NaNO_2$ in acidic conditions. (c) Azobenzene formation with a Tyr-bearing substrate.*

As the diamine is symmetrical, amide formation at either end results in the same product, while using it in excess reduces cross-linking of the resin beads, as it favours modification of a single amine. Aside from this adjustment, the diazotisation and S_EAr steps remained unchanged from those developed with SAv resin.

3.3.2 Catch-and-release using Affi-Gel

The catch-and-release performance of the Affi-Gel resin largely matched that of the SAv resin when considering small molecule and peptide substrates, with qualitative analysis by UHPLC and comparison with synthetic standards as before. In this instance, the modified peptide was sufficiently concentrated after desalting on a C_{18} flash cartridge to allow direct observation by ESI-TOF.

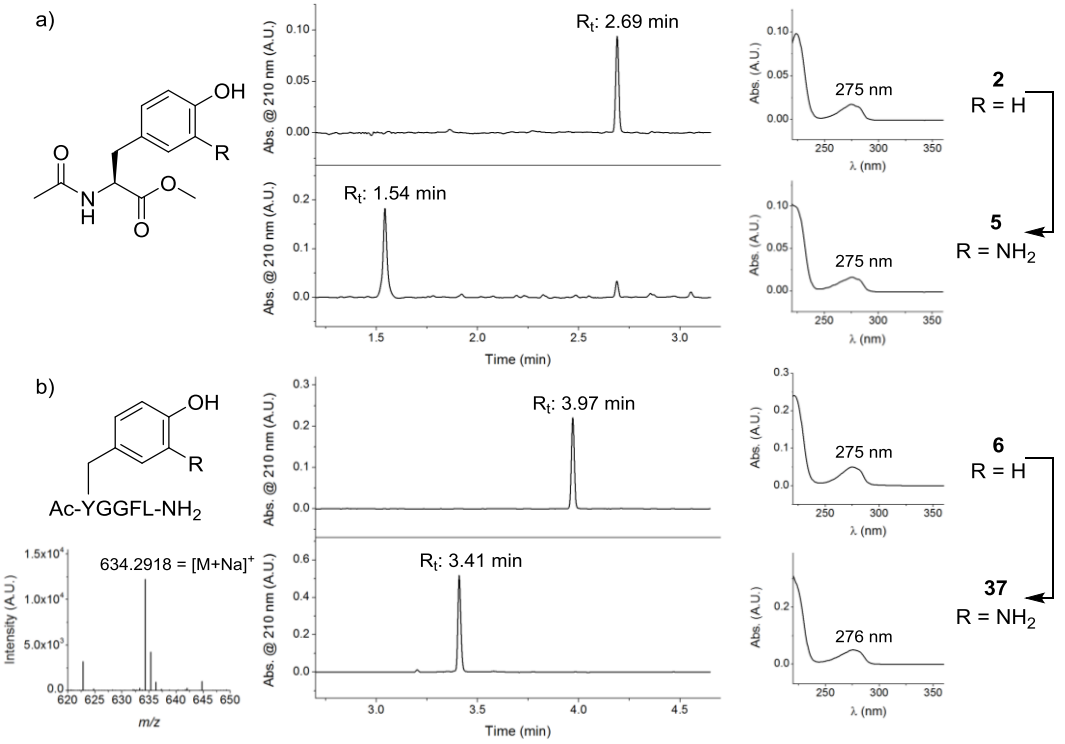


Figure 3-14 *Catch-and-release of small molecule and peptide substrates using Affi-Gel 10. (a) Ac-Tyr-OMe* ***2*** *was modified using the catch-and-release protocol. The product was identified by UHPLC with comparison to the standard* ***5****. (b) Leu-enkephalin* ***6*** *was modified in the same manner, the product was identified by UHPLC, with confirmation by ESI-MS.*

Just as with the SAv resin, the catch-and-release process was monitored by UHPLC and a shorter retention time was observed for each product compared to the respective starting material (Figure 3-14). With successful modification of a small molecule and peptide substrate, the Aff-Gel platform was next used for catch-and-release modification of Mb.

3.3.3 Catch-and-release of Mb

Affi-Gel resin was next used for catch-and-release of Mb, with analysis by UHPLC-MS. Compared to an unmodified Mb standard there was no significant shift in retention time upon catch-and-release modification of the protein (Figure 3-15). It was clear from the protein charge envelope that the modified sample differed from the standard, and this was confirmed following MaxEnt deconvolution, which converts multiple charge states to the deconvoluted uncharged mass (discussed further in Chapter 4).

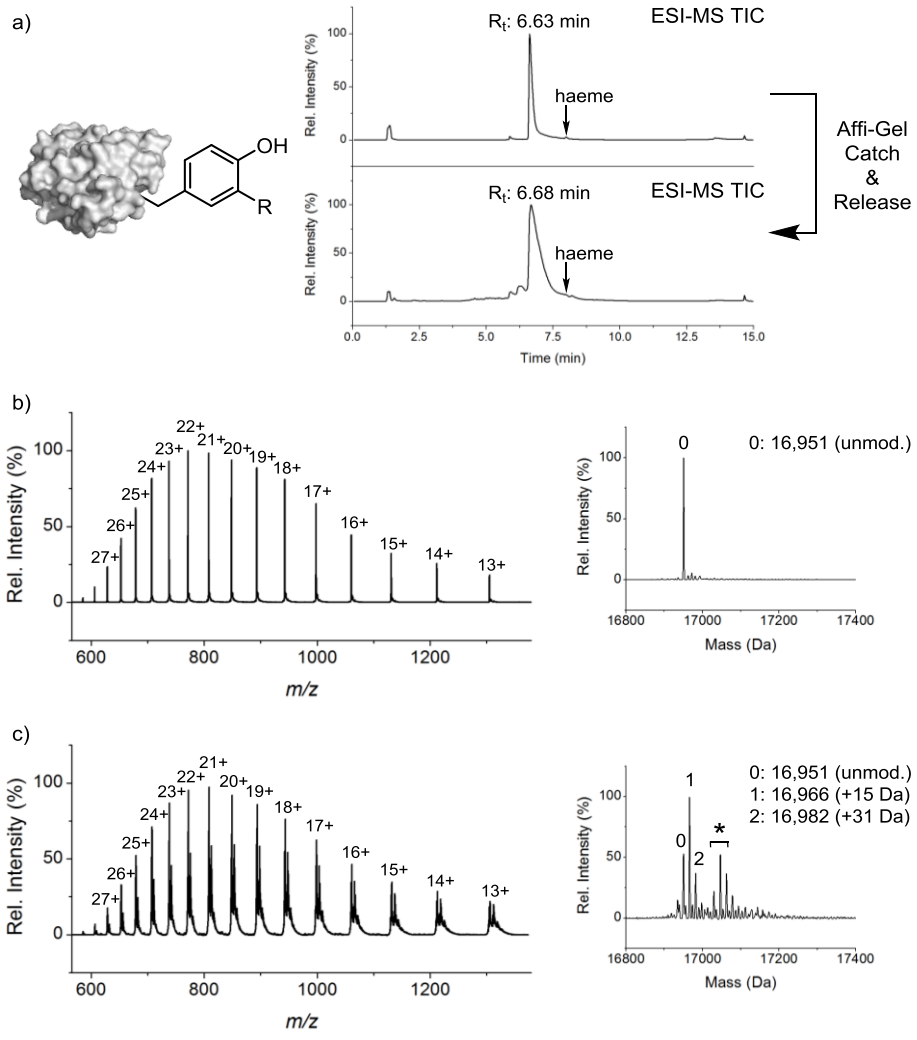


Figure 3-15 *Catch-and-release modification of Mb using Affi-Gel resin. (a) By LC-MS, modified Mb was identified at 6.68 min (Total Ion Chromatogram is shown). (b) ESI-MS of the unmodified Mb standard. The charge envelope of Mb is shown (left) alongside the MaxEnt deconvoluted spectrum (right). (c) The charge envelope and deconvoluted spectrum of modified Mb showed primarily singly modified Mb (1), alongside some unmodified (0) and twice modified protein (2). The same peaks were also identified bearing a phosphate adduct (*).*

The unmodified standard gave a mass of 16,951 Da, and it was gratifying to observe that the most prominent species in the modified sample showed a mass of 16,966 Da, corresponding to the +15 Da mass change indicative of the $-NH_2$ modification. Some unmodified and twice modified protein were also observed in this sample, alongside peaks displaying a phosphate adduct. This result represented a breakthrough in the catch-and-release project, as the first example of protein modification using an immobilised diazonium salt system. There remained some unanswered question as to the selectivity of the protocol, and this is explored further in Chapter 4.

3.4 Residue specificity

The residue specificity of the catch-and-release protocol was next explored, as there have been previous literature reports of diazonium salt reactivity with electron-rich aromatic amino acids other than Tyr, in particular Histidine (His)[141] and a hydroxyl-modified aromatic residue: 5-hydroxytryptophan.[142] In order to explore the residue specificity of the catch-and-release system, a peptide-based approach* was used to assess the likelihood of the immobilised diazonium salt showing cross-reactivity to amino acids other than Tyr. The anti-GFP (α-GFP) nanobody was identified as a potential future target for protein modification, as it bears much of the binding capability of a full-sized antibody, but at ~13 kDa, it is roughly one tenth of the size.[143] The combination of binding affinity and small size are key attributes for a protein-based probe, and so the α-GFP nanobody sequence was used as a template from which a useful peptide sequence could be selected, synthesised and modified.

3.4.1 *anti*-GFP nanobody sequence

Among seven Tyr residues in the nanobody sequence, Y081 was of particular interest (Figure 3-16). This residue is found on the opposite face from the binding region of GFP, and appeared to show high surface-accessibility. As a target for a peptide-based project, Y081 was also appealing as there are several polar residues nearby, which would aid the solubility of a synthetic peptide based on a Y081-containing sequence.

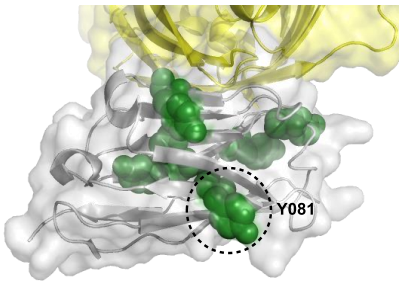


Figure 3-16 *The anti-GFP nanobody bound to GFP (PDB ID: 3OGO). The GFP protein is shown in yellow, Tyr residues in the nanobody are shown as green spheres. Y081 is highlighted as an attractive target, as it is not involved in the nanobody-GFP binding interaction, and appears highly surface-accessible.*

* Synthesis, purification and catch-and-release modification of the peptides was performed by project student Christina Burr; MS analysis was performed by CB & CA.

The native sequence which was selected for synthesis was ^{74}DDARNTVY^{81} (Figure 3-17), and this was synthesised by SPPS in a similar manner to Leu-enkephalin in Chapter 2. Five canonical amino acids were identified which may be expected to show some cross-reactivity with the diazonium salt, and these were substituted into the peptide sequence in place of A076, to give six synthetic peptides with the sequence Ac-DDXRNTVY-OMe. The 'X' amino acids comprised nucleophilic or aromatic residues which may act as competitive nucleophiles with the intended Tyr residue. Each peptide was modified using the catch-and-release protocol, before analysis of the peptide modification site.

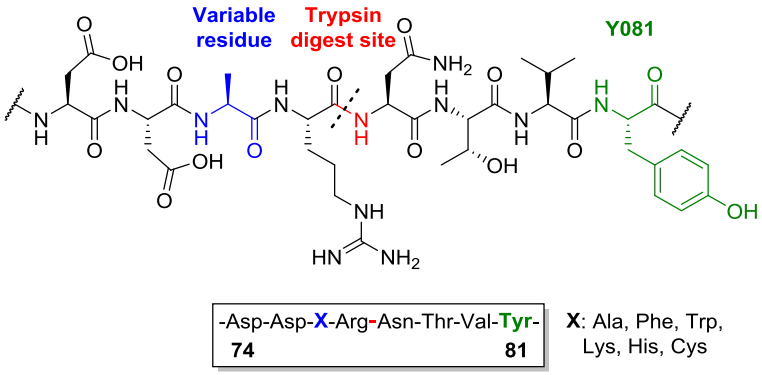


Figure 3-17 *An anti-GFP derived peptide sequence for catch-and-release modification. A peptide sequence containing Y081 was selected for synthesis, the Ala residue was identified as a site for substitution of other potentially reactive amino acids. The sequence also bears an Arg residue, which is a trypsin cleavage site.*

Initially, this project was conceived to allow assessment of the peptide-tagging site by UHPLC-MS, following enzymatic digestion of the tagged peptide. Using trypsin for the enzymatic digestion step would achieve cleavage at the C-terminal of R077 (Figure 3-17), effectively separating the peptide into two peptide fragments, each comprising four amino acids. UHPLC-MS of these digest products would then allow identification of the modification site on either the Tyr-containing or X-containing fragment. In practice, enzymatic digestion of the model peptide Ac-DDARNTVY-OMe gave rise to peptide fragments which did not retain well on the C_{18} column of the UHPLC. This did not bode well for the catch-and-release modified peptides, as these had an additional amine which would only serve to reduce the ability of the fragment to retain on a reverse-phase column. Instead, the decision was taken to analyse the intact catch-and-release modified peptides using top-down fragmentation by collision-induced dissociation (CID – discussed further in Chapter 4). The benefit of this approach over enzymatic digestion is the opportunity to achieve greater resolution of the site of modification, depending on the peptide sequence coverage achieved.

3.4.2 MS/MS of peptides

Following catch-and-release modification, the peptides were analysed by UHPLC-MS/MS. The intact masses were first analysed to identify the extent of modification for each peptide substrate. In most cases, the intact mass matched the theoretical peptide mass +15 Da, corresponding to a single modification, and these masses were isolated and fragmented under 20–50 V collision voltage (Figure 3-18). Sufficient coverage was obtained in each instance to identify Tyr as the site of modification, as expected.

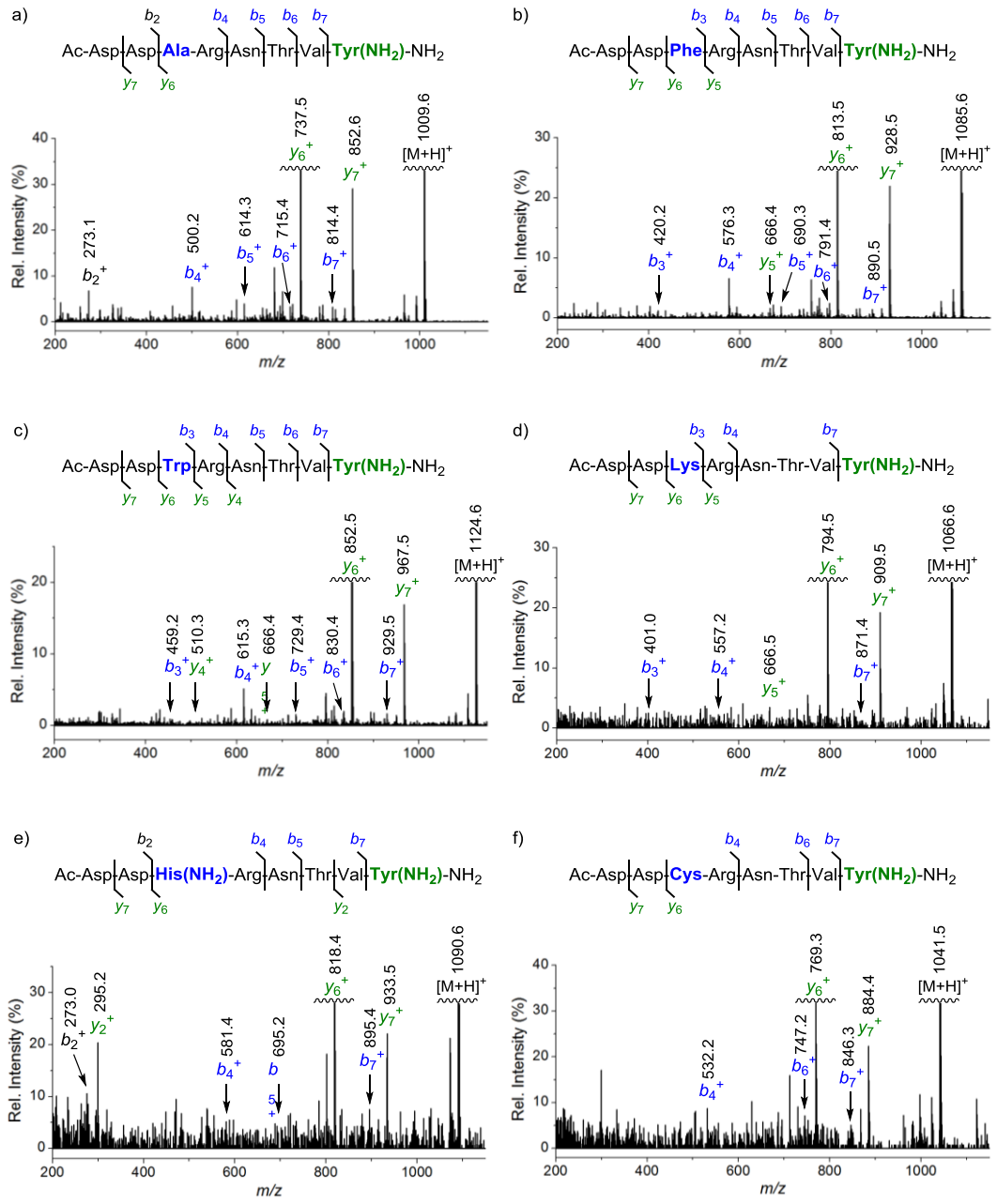


Figure 3-18 *Top-down fragmentation of catch-and-release modified peptides by CID.*

There was no evidence for cross-reactivity with Trp, Phe, Lys, Cys or Ala in the peptide screen, however in the case of Ac-DDHRNTVY-OMe, double modification was observed by LC-MS. Top-down fragmentation of this species revealed a +15 Da modification of the Tyr residue, as well as a +15 Da modification on the His-bearing end of the peptide [Figure 3-18 (e)]. Clearly, the catch-and-release system was not completely specific to Tyr-modification, and so future substrates would need to be analysed by top-down fragmentation to identify the precise location of modification, with particular attention to His residues as a potential site of cross-reactivity.

Histidine in fact shows even lower abundance in the proteome than tyrosine, and so while the peptide-based project had identified a potential pitfall in the application of the catch-and-release system, it was not considered sufficient to derail the project. In future tagging applications, careful analysis of the modified protein will instead be performed to identify cross-reactivity of any type.

3.5 Conclusion

In this chapter the catch-and-release chemistry which was developed in Chapter 2 was successfully transferred onto a solid support, initially through the use of SAv resin and a biotinylated aniline, and finally using the NHS-activated Affi-Gel 10 resin. The regiochemistry of the nitrosylation of biotin was explored following the unexpected synthesis of a nitrosobiotin derivative, and it was found that this modified substrate retains the ability to bind to SAv. Catch-and-release modification of small molecule and peptide substrates was performed, as well as successful modification of a protein substrate. Finally, the potential for cross-reactivity in the catch-and-release system was assessed and histidine was identified as a potentially competitive nucleophile for reaction with a diazonium salt. In future, protein substrates will be examined for the accessibility of histidine residues, and this will be considered as a potential modification site in post-cleavage analysis.

Chapter 4 Results & Discussion 3: Protein-tagging

4.1 *In silico* assessment of protein substrates

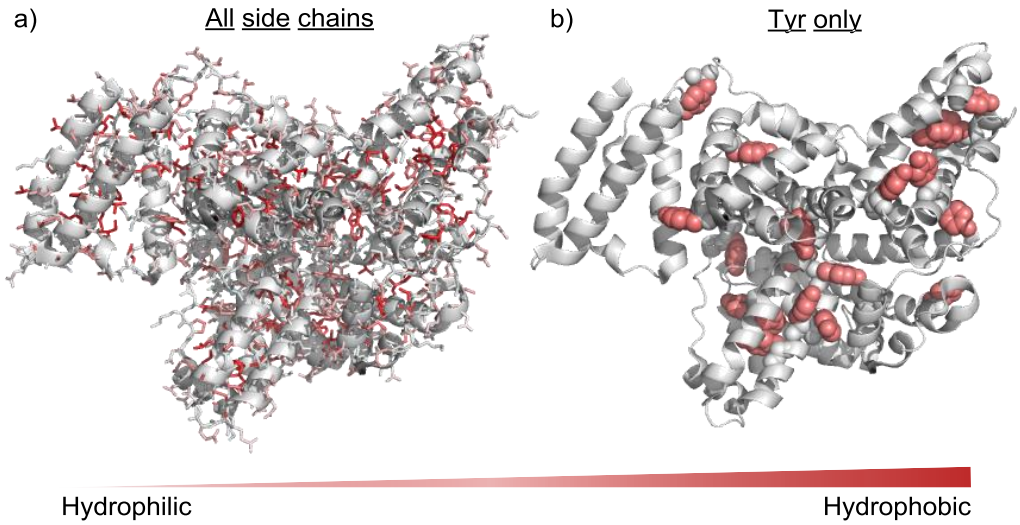
The application of the catch-and-release modification system was demonstrated in Chapter 3 using Tyr-containing substrates including a single amino acid, peptides and the model protein myoglobin. By inspection of the crystal structure, myoglobin contains only a single accessible tyrosine, and catch-and-release modification resulted in predominantly single modification. In order to further probe the selectivity of the solid-phase modification platform, other protein substrates were sought which bear multiple viable targets in the form of accessible Tyr residues. A further five substrates were selected [ribonuclease A (RNase A); soybean trypsin inhibitor (SBTI); β-lactoglobulin A (β-Lac); lysozyme C (HEWL); cytochrome C (Cyt. C)] which were commercially available, fell within the desired size limit (~10-20 kDa) and contained multiple Tyr residues.

In using a solid-phase platform for protein modification we expected to be able to achieve single modification of a protein substrate and to discriminate between multiple reactive sites based on surface-exposure. This can be justified by considering the dimensions of the agarose bead to which the Tyr-reactive diazonium salt is tethered, which has a reported diameter of 75–300 μm.[140] In the first instance, using an immobilised diazonium salt should favour modification at only the most exposed Tyr residues as the mobility of the electrophile is reduced, thus limiting access to more buried sites on the protein. Following modification, the protein is then covalently attached to the significant bulk of a resin bead, limiting the opportunity for secondary modification through a combination of steric effects and reduced diffusion of the immobilised protein through the solvent. The goal for this work was to exploit these effects for the development of a catch-and-release system which would provide single modification of a protein at a defined, predictable site, namely the most highly exposed Tyr. In order to assess the availability of Tyr residues in protein substrates, *in silico* methods were applied using the protein crystal structure, PyMOL software and the solvent-accessible surface area calculation.[144]

4.1.1 Tyrosine surface exposure

Beyond the primary structure of the polypeptide backbone, proteins exhibit higher order structure which is crucial to their function. Within a protein monomer, higher order structure is defined by a mixture of covalent and non-covalent intramolecular interactions. Covalent interactions which are important to protein structure include disulfide bridges between Cys residues, while non-covalent interactions may be polar (e.g. H-bonding) or non-polar (e.g. π-

stacking) in nature. In the resulting three-dimensional structure, a general trend is observed where amino acids with polar side chains are enriched on the protein surface where they may play a role in solubility and function [Figure 4-1 (a), grey side chains].[145] In contrast, non-polar amino acid side chains are most often found in the protein interior, where they are shielded from solvent and form hydrophobic pockets and clefts [Figure 4-1 (a), red side chains].[145]



c) Eisenberg Hydrophobicity Scale

Arg	Lys	Asp	Gln	Asn	Glu	His	Ser	Thr	Pro
-2.53	-1.50	-0.90	-0.85	-0.78	-0.74	-0.40	-0.18	-0.05	0.12
Tyr	Cys	Gly	Ala	Met	Trp	Leu	Val	Phe	Ile
0.26	0.29	0.48	0.62	0.64	0.81	1.06	1.08	1.19	1.38

Figure 4-1 *Amino acid hydrophobicity in bovine serum albumin (BSA) (PDB ID: 4J5S). Amino acid sidechains are coloured to represent their position on the Eisenberg hydrophobicity scale. (a) Hydrophobic amino acid side chains (red) tend to be found in the protein interior forming hydrophobic pockets. (b) Tyr residues (shown as spheres) display a range of surface-exposure across the protein sequence due to Tyr's ability to form both polar and non-polar interactions. (c) The Eisenberg hydrophobicity scale where Tyr ranks close to the mean of 0.*[31]

Due to its aromatic ring and polar head group, the Tyr side chain is considered amphipathic and scores close to the mean in Eisenberg's hydrophobicity scale (a consensus model used to rank the hydrophobicity of canonical amino acids).[31, 146] Tyr has the ability to form polar hydrogen bonds and also to participate in non-polar π-stacking interactions, leading to varied levels of surface exposure across the protein sequence [Figure 4-1 (b)].[147] Using *in silico* methods to interrogate the crystal structures of several proteins deposited in the PDB, a range of Tyr solvent-accessibility values have been observed, spanning from the highly surface-

exposed to those which were completely buried in the protein interior.[148] For this reason, Tyr can be considered an apt target with which to achieve selectivity based on surface-exposure.

4.1.2 The solvent-accessible surface area calculation

In order to rank the Tyr residues in a protein from most to least exposed, and therefore most to least likely to be modified in the catch-and-release system, a solvent-accessible surface area (SASA) calculation was performed using PyMOL software and the crystal structure of the protein of interest.[144] The SASA calculation was first proposed as a method by which to interrogate protein surface topology in 1971,[149] and is accessible as an embedded function in PyMOL.

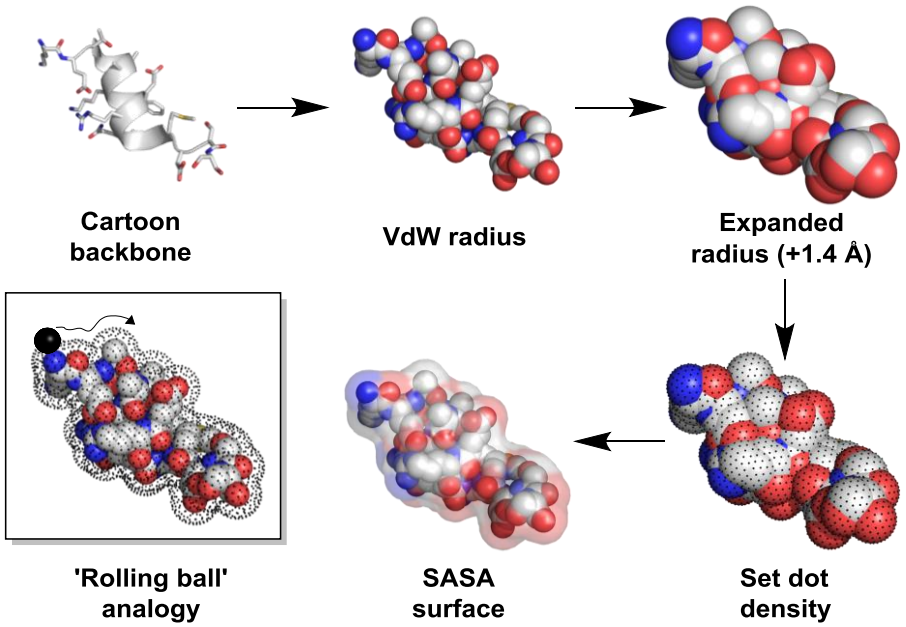


Figure 4-2 *The solvent-accessible surface area calculation. The Van der Waals (VdW) radius of each atom is first expanded by the radius of a solvent molecule. The surface coverage of data points is set by the 'dot density' and each dot is assigned as either 'exposed' or 'buried' depending on its state of overlap with a neighbouring atom. The 'exposed' data points are then used to calculate the SASA. A common analogy for the SASA calculation involves rolling a ball of solvent over the protein's surface. [The truncated peptide sequence shown is RNase A (1-16) (PDB ID: 4J5Z)].*

A common analogy used to explain the SASA calculation refers to a sphere of solvent rolling over the surface of the protein, with the output of the calculation representing the contact area between the solvent and the atom of interest (Figure 4-2). A high SASA value therefore represents a large solvent contact area, which is interpreted as a measure of high surface-exposure. In practice, to model the 'rolling' of a solvent molecule, the Van der Waals radius of each atom in the protein is expanded by 1.4 Å to approximate the interaction with a water

molecule.[150] Each data point on this expanded surface is then assessed by an algorithm which identifies regions where the expanded radii of neighbouring atoms overlap.[151, 152] The regions of overlap represent areas where solvent is excluded, and this is therefore subtracted from the total expanded surface area to give the solvent-accessible surface area with the units Å^2.

In order to execute the SASA calculation in PyMOL, the user first selects the atoms to be interrogated, in this case these are the reactive *ortho* positions of the Tyr phenol ring (named 'CE1' and 'CE2', following IUPAC abbreviation rules).[153] The user can perform the SASA calculation *via* the 'Action' menu in the Viewer Window [Figure 4-3 (a)], with a readout appearing in the External GUI. Alternatively, the same process can be performed using the Command Line in the GUI [Figure 4-3 (b)].

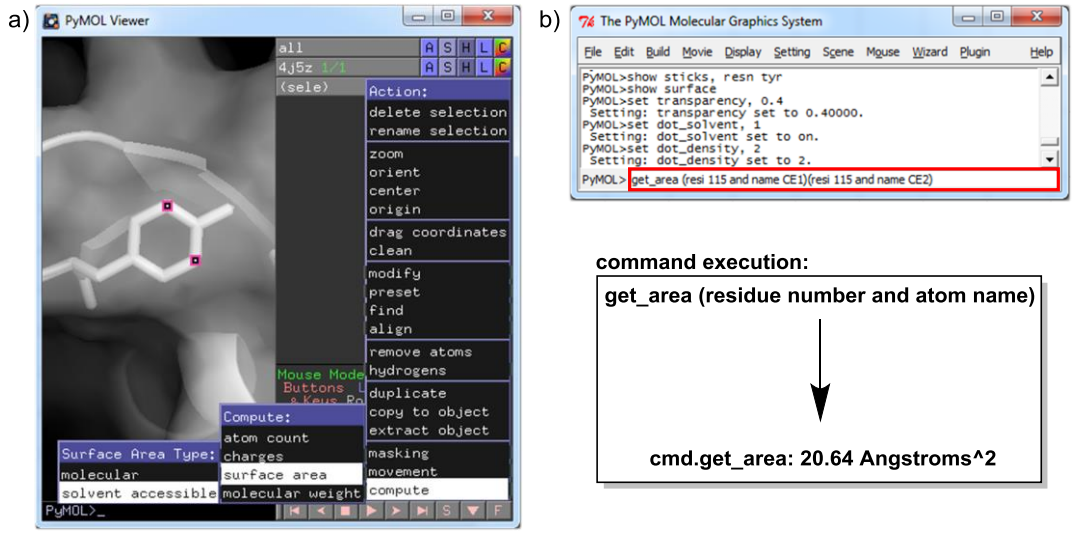


Figure 4-3 *Manual approaches to execute the SASA calculation. (a) The SASA calculation can be found in the drop-down 'Action' menu in the Viewer Window. (b) Alternatively, the SASA calculation can be executed by typing commands directly into the Command Line (highlighted in red) in the External GUI.*

A default solvent radius of 1.4 Å was used to probe the protein surface, modelling the solvent-accessibility of a water molecule. This variable could be altered to reflect the size of the residue-specific electrophile (in this case a diazonium salt). However, as demonstrated in the 'rolling ball' analogy, it is necessary to approximate the solvent probe to a spherical object of defined radius, and this becomes less realistic when modelling larger, linear molecules compared to a small solvent molecule such as water. In order to test the effect that increasing solvent radius had on the SASA output, Tyr residues in BSA were assessed using solvent radii of 1.4 Å, 1.9 Å and 2.4 Å (Figure 4-4). From these results it was apparent that using a large solvent radius reduced the sensitivity of the SASA calculation. In particular, residues with

intermediate (Y137, Y410) and low exposure (Y030, Y084) were barely distinguishable using a solvent radius of 2.4 Å, whereas the 1.4 Å radius calculation allowed comparison of these residues, whose SASA values were recorded as between 3.54-10.59 $Å^2$. A 1.4 Å radius was therefore used for future SASA calculations as it allowed evaluation of the highest number of Tyr residues, even those with intermediate or low exposure. It should be noted that the SASA calculation does not specifically model Tyr modification as it occurs in the catch-and-release system; rather, it provides an assessment of the relative exposure of each reactive site, which is taken as a predictor for the likely site of modification.

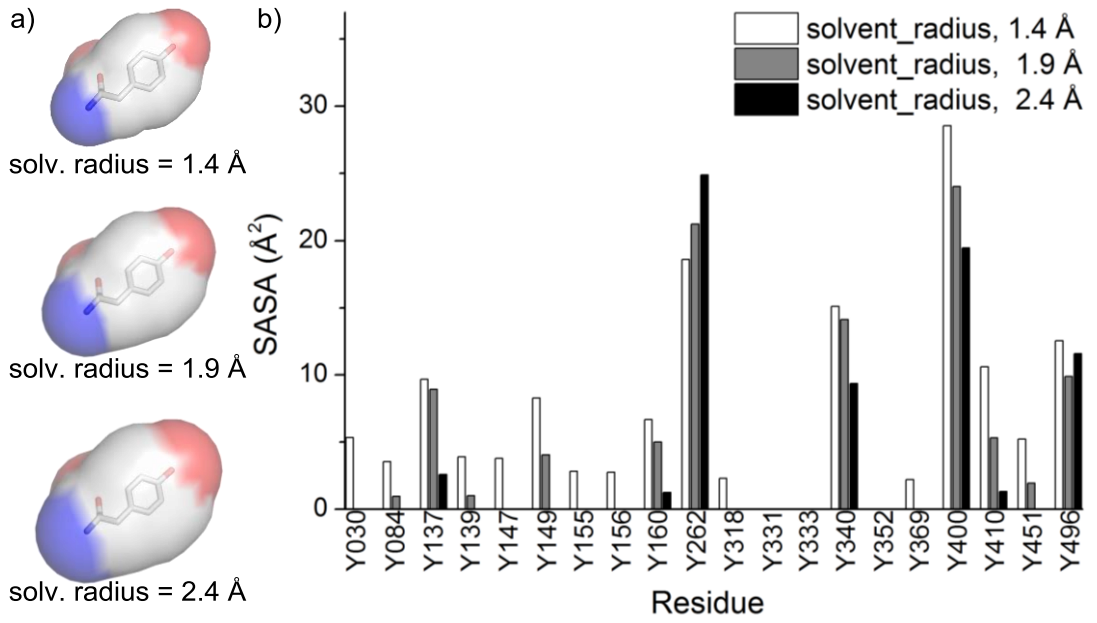


Figure 4-4 *The effect of increasing solvent radius on the SASA calculation. (a) The solvent radius variable defines the size of the expanded Van der Waals radius used (VdW surface shown for a single Tyr residue). (b) Three solvent radius settings (1.4 Å, 1.9 Å, 2.4 Å) were used for the SASA evaluation of Tyr residues in BSA (PDB ID: 4F5S). A solvent radius of 1.4 Å (white bars) gave superior resolution of intermediate and low exposure residues. All calculations were performed with a dot density of 2.*

Aside from the solvent radius, a second SASA variable to consider was the 'dot density', which can take a numerical value of 1-4 and dictates the density of data points on the expanded Van der Waals radius of the atom [Figure 4-5 (a)]. A high dot density gives a more accurate SASA output as there is more complete coverage of the surface and therefore more data points to report as 'buried' or 'exposed'. However, the cost of using a high dot density setting is the increased processing time required to perform the calculation.

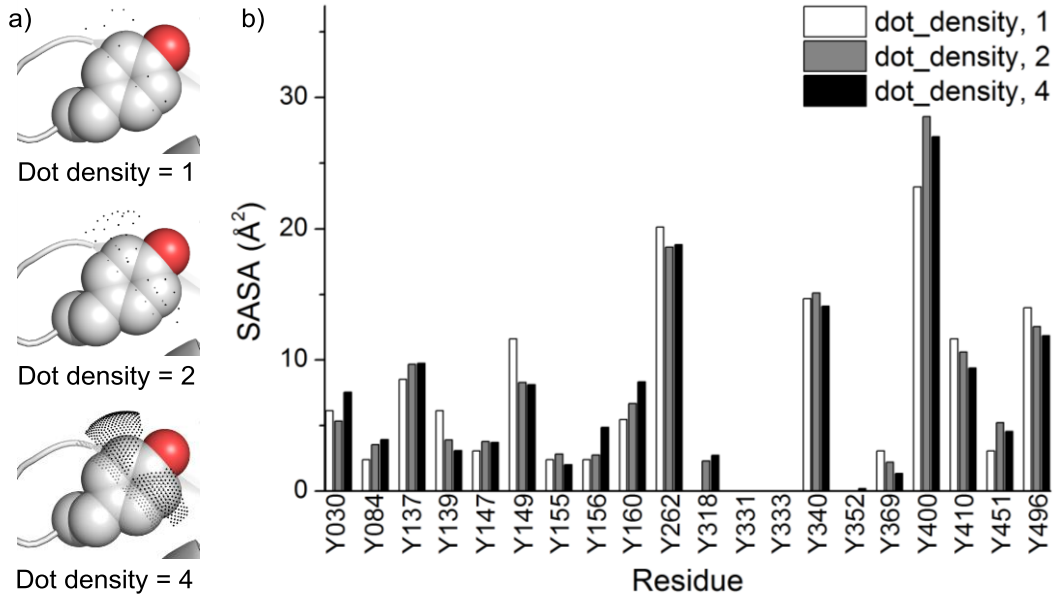


Figure 4-5 *The effect of dot density on the SASA calculation. (a) The dot density value dictates the number of data points on the modelled surface (shown here for the reactive ortho positions on a Tyr phenol ring).(b) The SASA of each Tyr residue in BSA (PDB ID: 4F5S) was calculated using three dot density (dd) settings. Total Tyr SASA and execution times were as follows: dd = 1 (137.82 Å², 41.71 s); dd = 2 (141.95 Å², 121.11 s); dd = 4 (141.32 Å², 1793.01 s). all calculations were performed with a solvent radius of 1.4 Å.*

The output of the SASA calculation was compared when using a low, medium or high dot density in order to find the optimal setting. Using BSA as an example, the SASA output of the twenty native sequence Tyr residues were found to follow a similar trend [Figure 4-5 (b)]. While the lowest accuracy setting (dd = 1) gave a total Tyr SASA which was comparatively low (137.82 Å^2), the total Tyr SASA of the medium (dd = 2) and high accuracy (dd = 4) settings were in reasonable agreement (141.95 Å^2 and 141.32 Å^2), providing evidence that a compromise between data point coverage and processing time could be struck without significantly affecting the reliability of the calculation.

Compared to the somewhat conserved total Tyr SASA values, the processing cost of each dot density setting was markedly different. The execution time for the SASA calculation of all reactive Tyr sites in BSA was recorded in PyMOL, and results ranged from 41.71 seconds for the low accuracy setting (dd = 1) to 1793.01 seconds for the highest accuracy setting (dd = 4). Although execution time will vary depending on factors including the number of Tyr residues in the protein, the number of chains in the crystal structure and on the processing capacity of the system used for the calculation, the highest accuracy dot density setting was considered too time-consuming for the current application. As such, the intermediate dot density setting (dd = 2) was selected as a good compromise between accuracy and processing time.

4.1.3 'Whole residue' versus 'reactive site' SASA

The SASA evaluation of endogenous amino acids in proteins has so far only considered the reactive positions of the residues of interest, in our case these are the *ortho* positions of the Tyr phenol side chain. Literature reports of similar methods to assess amino acid surface-exposure have not necessarily used this approach and have instead considered the exposure of the whole amino acid residue.[44, 45, 154]

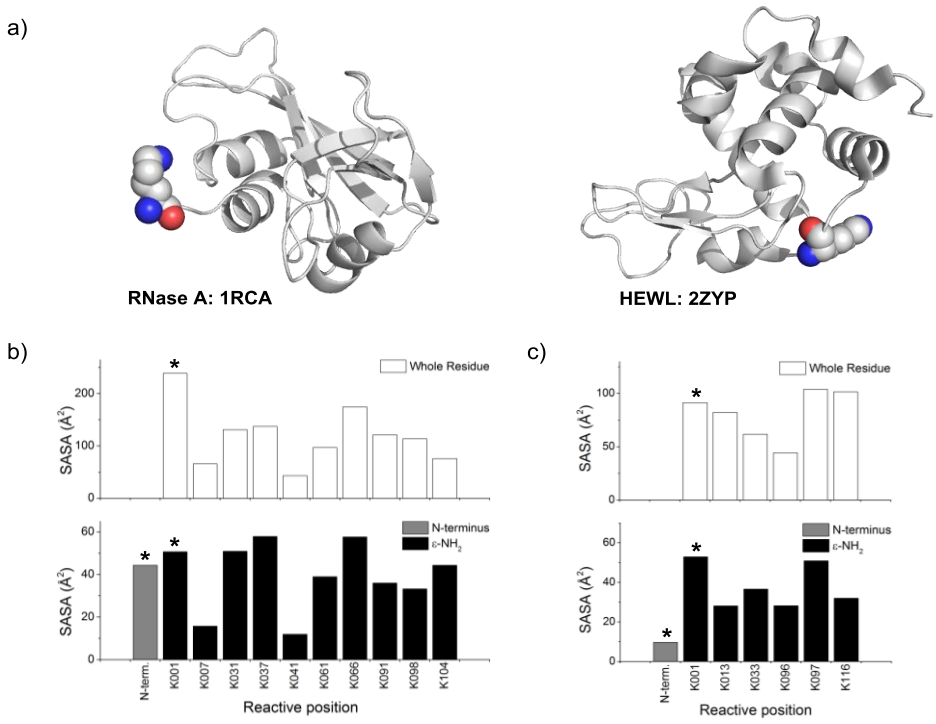


Figure 4-6 *Comparison of the 'whole residue' and 'reactive position' approaches to SASA calculations. (a) The structures of RNase A and HEWL used in the work by Weil et al.[44] In each instance the Lys residue at position 1 (K001, shown as spheres) was found to be modified. (b) The whole residue approach (top) and 'reactive site' approach (bottom) were compared for RNase A and (c) HEWL. Assessment of the N-terminus SASA is included in the 'reactive site' approach and modified residues are identified with *.*

These two approaches were compared using data reported by Weil *et al.*,[44] who achieved selective mono-labelling of a Lys residue in two native sequence proteins by treatment with sub-stoichiometric portions of an amine-reactive NHS-ester. As the targeted amino acid in this case is Lys, the reactive site in question is the nucleophilic ε-NH_2 of the Lys side chain. The proteins used in the study were RNase A and HEWL which both have a Lys residue at the N-terminus. In each case, the modified residue was identified by MS/MS as K001 [Figure 4-6 (a)] and this was justified using a 'whole residue' solvent-accessibility profile of the protein in question. Using the same crystal structures as the published work, the proteins were assessed for SASA on a 'whole residue' and 'reactive site' basis [Figure 4-6 (b) and (c)]. The published

work reported that K001 in RNase A was found to have the highest solvent-accessibility and this was supported by the 'whole residue' SASA analysis. However, by the 'reactive site' method, K001 was only the fourth most exposed reactive site behind K060, K037 and K031. In addition, the N-terminus was found to have high exposure and therefore this site should also be considered as a potential labelling position. In fact, given the reactivity of the N-terminal α-amine (whose pKa is around 2 units lower than the Lys ε-NH_2),[48] it is quite possible that this was the true location of modification in the reported work. The authors identified the location of modification by MS/MS, however this would not allow distinction between the N-terminus and the side chain ε-NH_2 of the same residue. HEWL was report to be modified at K001 despite this residue ranking as the third most exposed Lys by the 'whole residue' method.[44] However, the 'reactive site' approach identified K001 as the most exposed Lys ε-NH_2, which would better explain the findings. In this case the N-terminus was not highly exposed, but modification at this position should not be ruled out due to the higher reactivity of this amine.

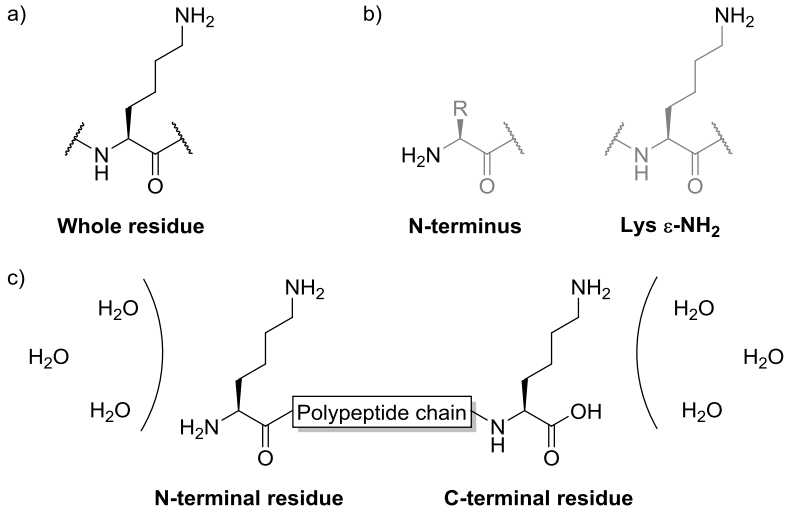


Figure 4-7 *The contributing atoms in a Lys residue when using the 'whole residue' and 'reactive site' SASA calculations. (a) The 'whole residue' approach considers the exposure of non-reactive atoms in the amino acid as well as the reactive ε-NH_2. (b) The 'reactive site' approach considers only the reactive atoms and allows separate assessment of the N-terminus and the Lys ε-NH_2. (c) The 'whole residue' approach is likely to overstate the exposure of N- and C-terminal residues as they are attached to the polypeptide chain on one side only.*

The primary concern about using a 'whole residue' approach is its consideration of non-reactive atoms in the amino acid (Figure 4-7). For a mid-chain Lys residue this means that atoms in the peptide backbone as well as the carbon atoms in the side chain make a contribution to the 'whole residue' SASA value, despite the fact that they are not reactive to an NHS-ester. Compared to a mid-sequence residue, the 'whole residue' approach is also likely to overstate the solvent-accessibility of an N- or C-terminal amino acid because it is attached to the

polypeptide chain on one side only. In this case the backbone and non-reactive side chain atoms of a terminal residue are likely to be more exposed to solvent than in an amino acid which is flanked by residues on either side [Figure 4-7 (c)]. Finally, in taking a 'whole residue' approach there is no means to separate the reactive N-terminus from the side chain of the N-terminal residue, and this is especially problematic when using an NHS-ester for protein-tagging as the N-terminus is more nucleophilic than the Lys ε–NH_2.[48] For these reasons, the 'reactive site' approach was preferred for SASA assessment of amino acids in this project.

4.1.4 Automation of the SASA calculation

Manual control of the SASA calculation had so far proved useful for interrogating single targets but this was ultimately an inefficient way to screen a large number of residues. To expedite the process, the task was automated by the development of a program which can run in PyMOL and export the results of the SASA calculation into a .txt file for easy analysis and presentation.* The script was written in the PyMOL-compatible language Python to send the necessary commands to perform the SASA calculation. This was primarily focussed on Tyr as the native sequence target, but it has also been expanded to assess the exposure of other common endogenous targets (Lys, Cys, and the α-NH_2 of the protein N-terminus) as well as His to assess the potential for cross-reactivity. For the sake of clarity, the following description of the SASA automation script will consider the Tyr target only, as each amino acid variant uses a similar structure.

The automated SASA script largely follows the steps which one would take to perform the calculation manually (Table 4-1). As PyMOL has a well-defined amino acid and atomic naming system, the initial selection of the reactive *ortho* positions on the Tyr phenol ring involved sending the command "select (resn Tyr and name CE1)(resn Tyr and name CE2)". However, selection of all reactive positions at once does not allow the distinction of separate Tyr residues, and so a list of delineated reactive positions was created to allow SASA evaluation of single atoms. This task involved separating the atoms using all the identifying information available (PDB number, chain identifier, residue number, amino acid name, atom name). The list was then used to populate a dictionary to which the SASA output of each atom could be attached (in Python the atom identifiers are termed 'keys', and the SASA output 'values'). Before execution of the SASA calculation, the 'dot_solvent' and 'dot_density' variables were set as described previously. The SASA calculation of each reactive site was

* The development of the automated SASA script was performed by project student Mirek Kosar, expansion of the script to include SASA analysis of Lys and the N-terminus was performed by CA.

executed using the "get_area" command, and the resulting values were added to their respective keys to give a complete dictionary. As there are two reactive sites on the Tyr ring ('CE1' and 'CE2'), the SASA output of each atom can be given separately or as a summed 'SASA per residue' value.

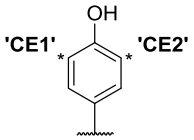
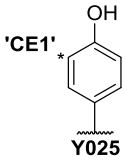
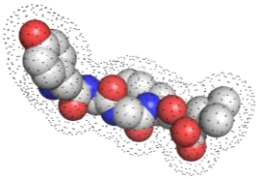
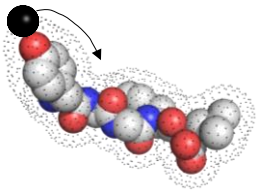
Operational Step	Key Commands	Graphical Annotation
Select reactive positions	`target_sel = (resn, name)`	
Delineate selected atoms	`list.append(pdb, chain, resi, resn, name)`	
Set solvent radius, dot density	`cmd.set('solvent_radius', 1.4)` `cmd.set('dot_density', 2)`	
Calculate SASA of each selected atom	`cmd.get_area(resi, name)`	
Output SASA to .txt file	`f.write('Residue: %s, SASA: %f')`	Residue / SASA(Å²): Y001 3.02; Y002 7.76; Y003 29.36

Table 4-1 *An outline of the key steps in the automated SASA calculation program. To overcome the laborious process of manually selecting each reactive position, the script constructs lists and dictionaries to identify and separate these sites, then collates the SASA output in a .txt file. The peptide sequence shown is Leu-Enkephalin (PDB ID: 5E3A).* ***Key****: resn = amino acid name; name = atom name; pdb = PDB ID; chain = chain number; resi = residue number.*

For the current application the summed 'SASA per residue' output was preferred as this allows the user to rank each amino acid based on the surface-exposure of the combined reactive positions. Further to this, it would be unlikely that the final protein modification could be located to a single atom using the available analytical techniques, so separating the two *ortho* positions of a single residue was of no practical use. To convert the SASA results into a format which was easy to handle, a blank .txt file named 'tyr_sasa_out.txt' was created by the Python command 'open('tyr_sasa_out.txt', 'w')'. Into this file was written the atom identifiers and their associated SASA output. From here, the SASA results may be transferred to graph-plotting software such as Excel or Origin, giving a straightforward method to evaluate Tyr exposure by residue, and assess the target protein's suitability for the catch-and-release system.

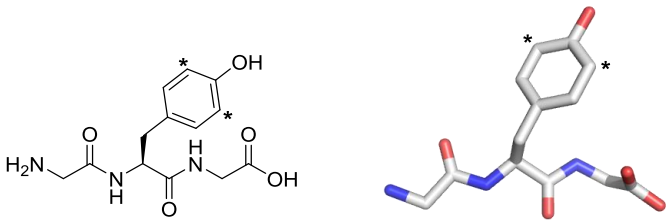


Figure 4-8 *Tripeptide GYG (2-D representation,* ***left****) was used to calculate a benchmark SASA value for the ortho-positions of the Tyr phenol ring. The peptide structure was drawn in ChemSketch and 'cleaned' in PyMOL to give the optimised 3-D structure (****right****).*

In order to better understand the SASA output, the calculation was performed on a tripeptide in which a Tyr residue was flanked by two Glycine residues. The GYG peptide is the simplest representation of a mid-chain Tyr residue as the Gly residues on either side bear no side chains which would affect Tyr accessibility. This model has been used as a benchmark of high solvent-accessibility in other formats including the Molecular Operating Environment software used by Weil *et al.*[44, 155] The molecule was drawn using ChemSketch software (Figure 4-8),[156] and transferred into PyMOL where the 'clean' operation was used to refine the 3-D structure by optimising bond angle, bond length, H-bonding and electrostatic effects.[157] The SASA command was then executed to give an output of 23.10 Å^2 against which other Tyr residues could be benchmarked.

4.2 Protein substrates for catch-and release modification

The automated SASA script was used to assess protein candidates for catch-and-release modification. A desirable protein substrate would have multiple Tyr residues with varied SASA values to allow discrimination based on surface-exposure. In addition, the protein should be of a similar size to future nanobody or other small protein substrates (~10–20 kDa) and should be soluble at ~pH 7. Following the observation of cross-reactivity with His residues in peptides, the SASA scripts include the calculation of solvent-accessibility of His residues. None of the candidate proteins were discounted due to their His SASA results, but the values were noted in case of future discovery of His modification by MS/MS. Six candidate proteins were selected (Figure 4-9) with a range of properties including M_W (12–21 kDa), pI (4.5–12), as well as the number (2–6) and exposure of Tyr residues (0–38 Å^2).

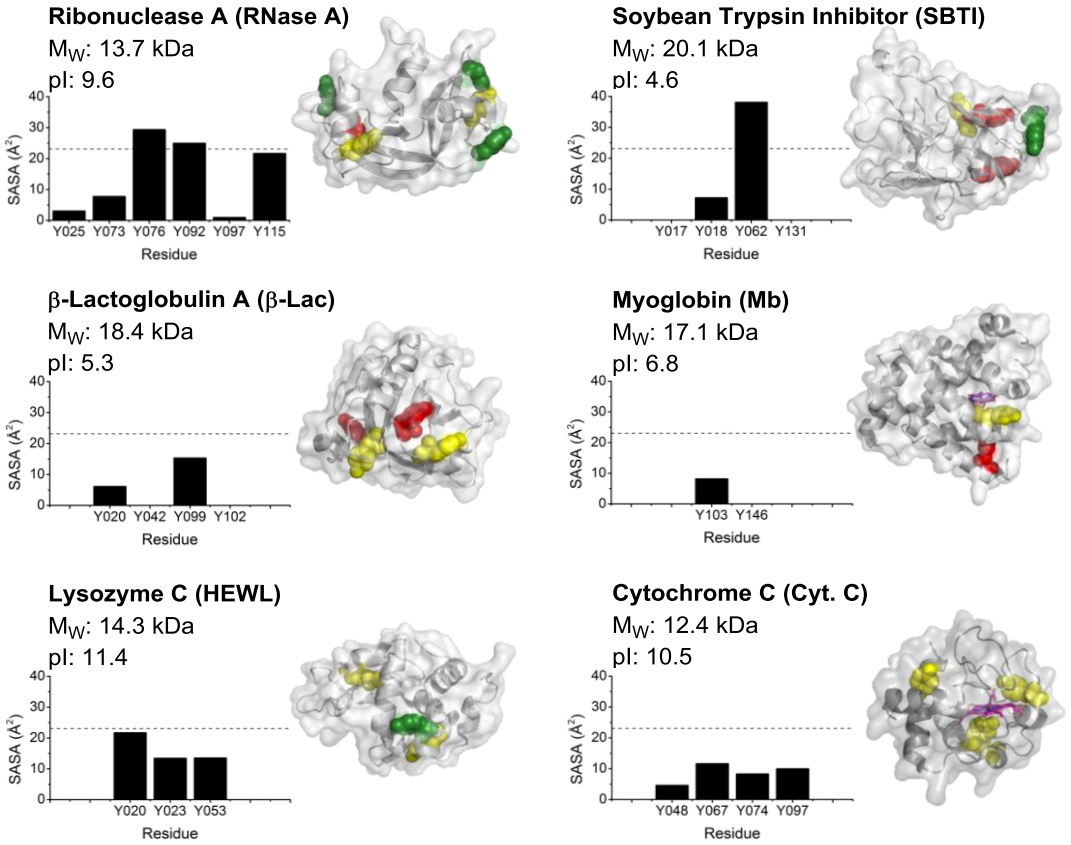


Figure 4-9 *Assessment of Tyr SASA in six protein substrates. Tyr residues which are highly exposed (SASA >70% of GYG tripeptide) are coloured green, intermediately exposed residues are coloured yellow (10-70%), and the least exposed Tyrs are coloured red (<10%). The haeme groups in Mb and Cyt. C are shown in purple. The GYG peptide benchmark is shown as a dashed line at 23.10 Å^2. (PDB ID's: 3NPO; 2B4Z; 4YM8; 5D5R; 4J5Z; 1BA7).*

4.2.1 Ribonuclease A as a catch-and-release candidate

Of the proteins considered, RNase A was highlighted as a particularly interesting substrate as it has six Tyr residues in the native protein sequence with a range of SASA values [Figure 4-10 (a)]. There were three Tyr residues with high exposure (SASA: >20 Å^2), one of intermediate exposure (SASA: ~8 Å^2), and two Tyrs which were mostly buried (SASA: <5 Å^2). This analysis suggested that it would be possible to selectively target the more exposed Tyrs in the protein sequence (Y076, Y092 and Y115, coloured green in Figure 4-10) without labelling the others (Y025, Y073, Y097). As RNase A has three highly exposed Tyrs, it was also considered an interesting target to test the hypothesis that the solid-phase catch-and-release system would favour single modification of the protein, even when other reactive sites were available and highly exposed.

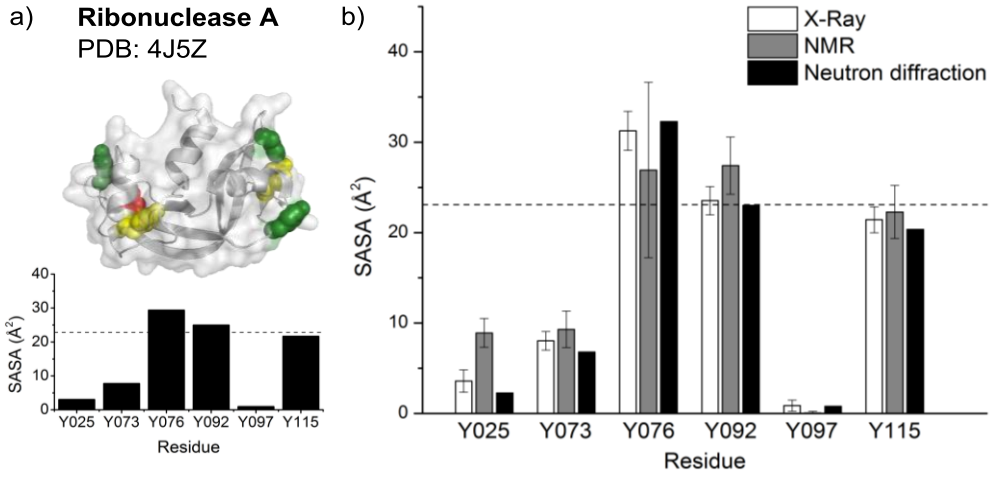


Figure 4-10 *The conserved solvent-accessibility profile of native sequence RNase A structures. (a) The crystal structure used for SASA evaluation of Tyr residues in RNase A (PDB ID: 4J5Z). (b) X-ray crystal structures of native sequence RNase A were retrieved from the PDB (n = 75). The average of these results was compared with an NMR solution structure (average of 32 conformers is shown) and a neutron diffraction structure, and all showed a similar solvent-accessibility profile. The GYG peptide-derived benchmark is shown as a dashed line.*

RNase A is a well-studied protein, with >100 X-ray structures deposited in the PDB. Of these deposits, 75 examples were found of the native sequence protein and these were evaluated for Tyr SASA [Figure 4-10 (b)]. The solvent-accessibility profile was found to be well conserved among the X-ray crystal structures, and the average of this data set matched well with the structure previously used. While X-ray crystal structures were the most abundant RNase A deposits in the PDB archive, there were also structures which had been solved using NMR and neutron diffraction. The NMR structure was particularly interesting as this shows the conformation of the protein in solution rather than a solid-phase crystal structure. Between the

32 conformers which make up the NMR-derived deposit, the standard deviation was significantly higher than that observed between the entire set of X-ray derived structures, reflecting the dynamic nature of a protein in solution. Despite this variability, the average NMR-derived solvent-accessibility profile was in reasonable agreement with the average X-ray derived profile. The neutron diffraction structure was also in excellent agreement with the X-ray derived profile, providing evidence that the crystal structure used was representative of the wider structural evidence, and that this agrees with the conformation of the protein in solution.

4.3 Optimisation of the catch-and-release protocol

4.3.1 Analysis and desalting of a protein sample

In moving from small molecules and peptides to proteins, it was necessary to adapt the catch-and-release methodology due to the limited availability and increase in complexity of the protein substrates. In particular, analysis of the catch-and-release product was more difficult with a protein target compared to a small molecule. At each key step of the catch-and-release reaction, UHPLC analysis of the resin eluate was used to monitor progress. When using small molecule and peptide substrates, analysis by UHPLC using a C_{18} column allowed distinction of unmodified and modified substrates by their respective retention times. However, with protein substrates it was necessary to use a C_4 column for analysis and the addition of a single amino group proved to be an insufficient modification to cause a change in protein retention time. As such, the 'wash' step (coming after protein conjugation to the solid-phase and before reductive cleavage) was especially crucial for separation of unmodified and modified substrate. The wash eluate was therefore monitored by UHPLC until there was complete disappearance of the protein peak (Figure 4-11).

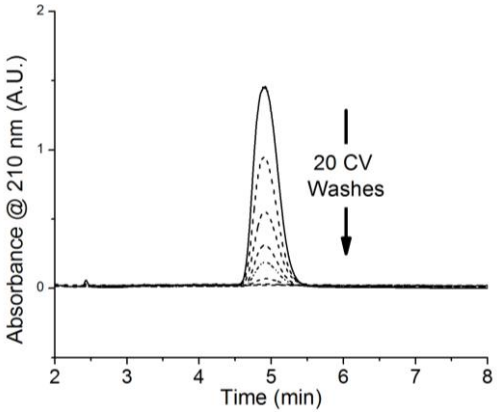


Figure 4-11 *The disappearance of unmodified RNase A in the resin eluate over 20 column volume (CV) washes.*

As the presence of modified protein could not be identified by a shift in retention time, it was now monitored by the disappearance of the protein peak in the 'wash' eluate, followed by reappearance in the 'cleavage' eluate. The samples were later characterised by MS to confirm the extent of protein-tagging, however the UHPLC was used for 'real time' reaction monitoring.

The catch-and-release cleavage product is collected in 300 mM $Na_2S_2O_4$ and this sample must be desalted to remove the reducing agent before the following oxidative coupling reaction can occur. For small molecules, a C_{18} solid-phase extraction column was used for this purpose, however a protein substrate would bind to this media and could not be eluted in the standard MeOH or MeCN without becoming denatured. More common methods for protein desalting include dialysis, size-exclusion chromatography (SEC) and passage through a molecular-weight cut-off (MWCO) membrane.

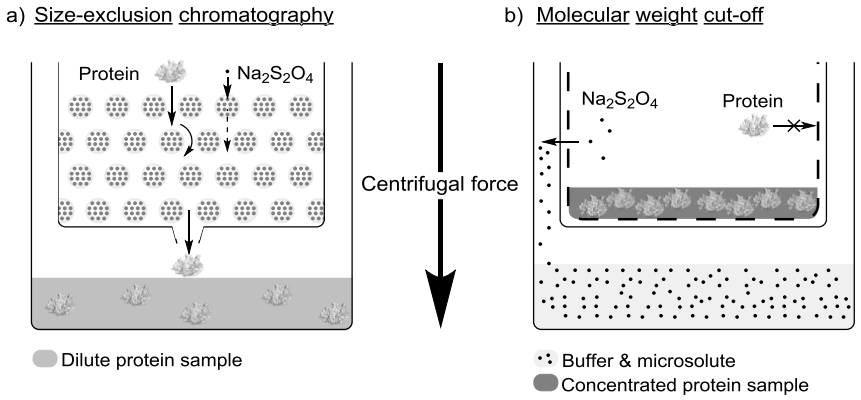


Figure 4-12 *The workup of the protein catch-and-release reaction. (a) SEC: using a size-exclusion spin cartridge, the passage of the salt through the resin bed is hindered by diffusion into the resin's pores. The macromolecule is too large to enter the pores and passes more quickly through the resin. (b) MWCO: the sample is desalted by successive concentration and dilution with buffer, using a membrane which allows the passage of microsolutes only.*

Desalting method	Protein recovery from H_2O (%)	Protein recovery from $Na_2S_2O_4$ (%)
SEC (100 µL)	95.0	8.1
MWCO (500 µL)	91.0	54.1

Table 4-2 *Table: Comparison of protein recovery using two desalting methods. No significant difference in recovery was observed for an RNase A standard in water, but the MWCO method gave significantly better recovery of RNase A from a 300 mM $Na_2S_2O_4$ solution.*

The use of dialysis was initially discounted as it is more commonly used on higher volume and higher concentration samples than the catch-and-release reaction produces. Instead, commercially-available SEC and MWCO spin cartridges were compared, using a centrifuge to pass the sample through the chosen media. Both approaches gave high protein recovery when trialled on a protein standard in water. However, when a sample containing 300 mM $Na_2S_2O_4$ was used there was a significant loss of protein in each instance, however the MWCO cartridge outperformed SEC in terms of protein recovery and so this method was adopted for sample desalting.

Despite these adaptations, early attempts at protein catch-and-release modification were extremely low yielding, reflecting the reduced accessibility of the diazonium salt to a Tyr residue in a protein as compared to a small molecule or peptide. This phenomenon is due in part to the reduced surface-exposure of Tyr residues in proteins, however the nature of the Affi-Gel resin may also be a contributing factor to this. The resin used in the Affi-Gel 10 product is a cross-linked polymer more often used for size-exclusion chromatography. Molecules below the exclusion limit (5,000,000 Da) can pass through pores in the media, gaining access to any activated sites which are buried within the internal volume of the resin bead. While the proteins under consideration have a M_W below the exclusion limit, their diffusion into the resin interior will be slower than that of a small molecule or peptide, restricting access to the activated sites. It was hoped that restricted access to the active sites would enhance the selectivity of the catch-and-release system, and so other variables were considered to optimise the protocol and increase the yield of the reaction.

4.3.2 Electronic effects in diazo chemistry

Following literature precedent, it was found that both the 'catch' and 'release' steps of the catch-and-release system could be optimised by reducing the electron density of the resin-bound aromatic ring [Figure 4-13 (a)]. It has been shown that the introduction of an electron-withdrawing group (EWG) to the aromatic ring of a diazonium salt serves to increase the stability of the diazonium cation.[158] In a system which relies on limited accessibility for selectivity, increasing the stability of the diazonium ion should serve to increase the window of time in which azobenzene formation can occur, thus improving the 'catch' step of the modification system. In order to improve the 'release' step, the azobenzene group should allow efficient reductive cleavage by $Na_2S_2O_4$. For the design of a highly labile azobenzene moiety, it has been reported that there should be a large disparity in electron density between the aryl rings on either side of the diazo bond.[159] In the current work, Tyr fulfils the role of an electron-rich arene, with an electron donating hydroxyl group in the *ortho* position to the reactive site. Reducing the electron density of the opposing aromatic ring should therefore result in a highly labile azobenzene construct. In this way, a single substitution on the resin-bound aryl ring should improve the overall yield of the catch-and-release system through simultaneous improvement of both the 'catch' and 'release' steps.

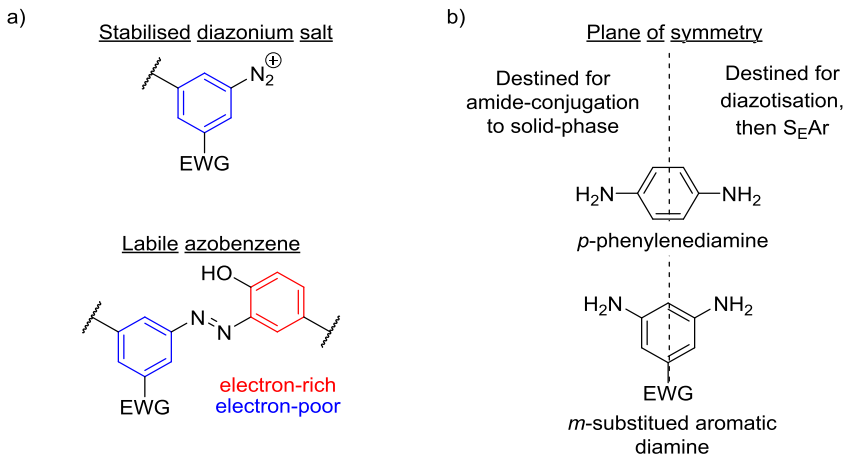
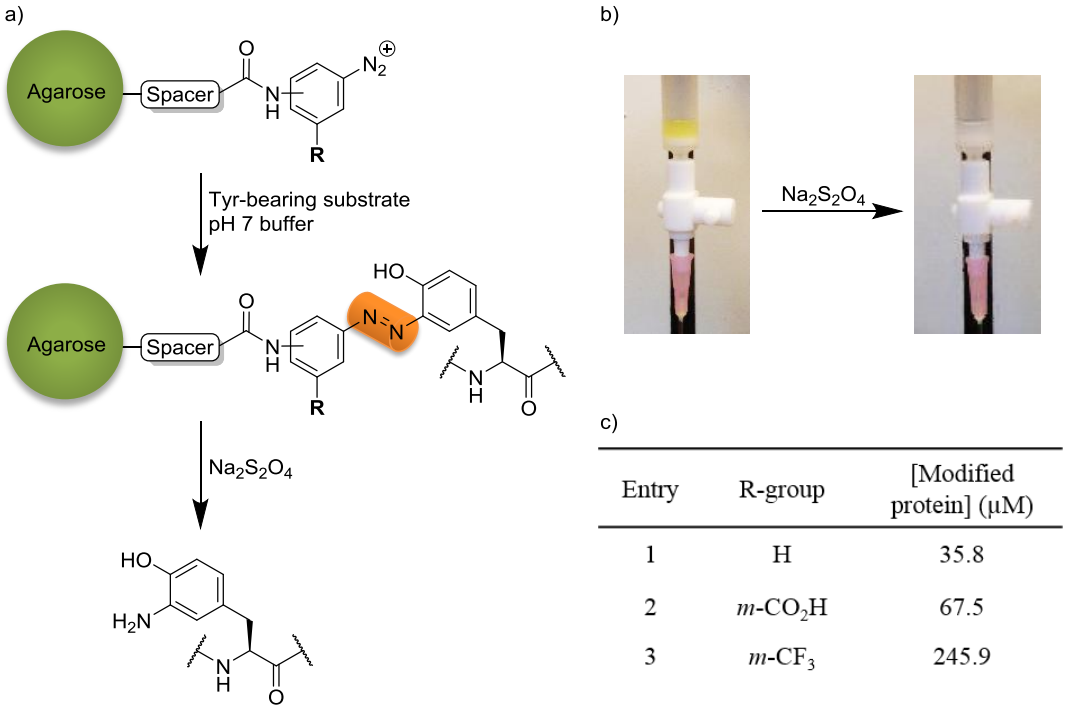


Figure 4-13 *Optimising the catch and release system through electronic effects. (a) Electronic effects on the stability of a diazonium salt and the optimal arrangement for a highly labile azobenzene moiety. (b) The importance of symmetry in the aromatic diamines used for this work.*

To preserve the one-step functionalisation of the resin, symmetrical diamines were preferred, leading to the consideration of *meta*-substituted aromatic diamines. When used in excess these reagents allow amide conjugation to the resin, while leaving an immobilised aniline for diazotisation, thus negating the need for amine protecting groups [Figure 4-13 (b)].

Aromatic diamines with -CO_2H and -CF_3 substitutions in the *meta* position were used in the catch-and-release protocol for the modification of RNase A [Figure 4-14 (a)]. In comparison to *p*-phenylenediamine, both electron-poor diamines gave an improved yield of the cleavage product. In agreement with the literature-based prediction, the more highly EWG (-CF_3) gave the highest yield, representing a nearly 7-fold improvement when compared with *p*-phenylenediamine.



Entry	R-group	[Modified protein] (μM)
1	H	35.8
2	*m*-CO_2H	67.5
3	*m*-CF_3	245.9

Figure 4-14 *Optimisation of the catch-and-release of RNase A using different diamines. (a) By switching from R=H to R=EWG, the diazonium salt cation is stabilised. (b) Cleavage of the azobenzene linker is accompanied by a loss of colour in the resin. (c) The yield of modified protein was increased by adding an EWG to the diamine, the most electron-withdrawing group (-CF_3) gave the highest yielding reaction.*

4.3.3 Reaction time and the persistent diazonium ion

Using the optimised system, a significant increase in yield was also observed with increasing reaction time. The initial conditions for catch-and-release modification required 24 h incubation of the protein with the immobilised diazonium salt, although it was suspected that the diazonium salt would either react or become quenched long before this time point. In particular, hydrolysis of the diazonium salt to give the corresponding phenol was considered the most likely route of degradation given the use of aqueous buffer as solvent in this reaction.

The displacement of a diazonium salt by a nucleophile is entropically favoured due to the production of $N_{2(g)}$ and this driving force is exploited in substitution reactions in organic chemistry, notably in named reactions such as the Sandmeyer and Balz-Schiemann reactions. A similar transformation can occur in an aqueous environment with a water molecule acting as the nucleophile, resulting in hydrolysis of the diazonium salt and production of a phenolic product. The rate of diazonium salt hydrolysis is known to increase with increasing temperature and pH, and this places the degradation route in direct competition with the desired S_EAr reaction, which is also accelerated in basic conditions (Figure 4-15).

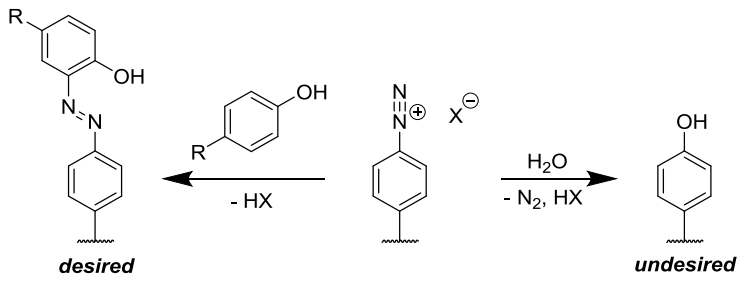


Figure 4-15 *Hydrolysis of the diazonium salt (right) may be a competitive and undesired side reaction when attempting S_EAr (left) in basic aqueous conditions.*

The catch-and-release reaction was performed at room temperature and pH 7.2 for optimum compatibility with protein substrates. However, it is likely that these neutral conditions also limit the rate of the undesired diazonium hydrolysis reaction, thus stabilising the diazonium ion for sufficient time to allow the desired S_EAr reaction to take place. In fact, it was observed that the yield of the S_EAr reaction with RNase A continued to increase with reaction time >24 h, reaching a maximum of ~70% at 48 h. This suggests that the reactive diazonium species remains unquenched for at least 24 h, and this is assumed to be a combined effect of the stabilising ability of the EWG on the diazonium cation and the mild temperature and neutral pH used for the reaction.

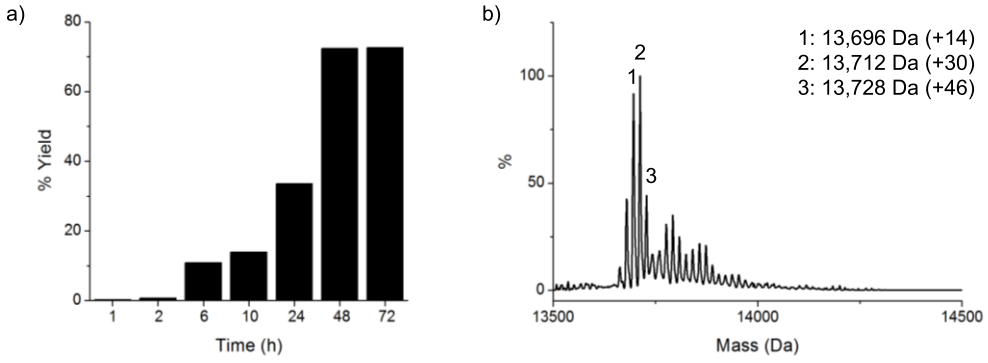


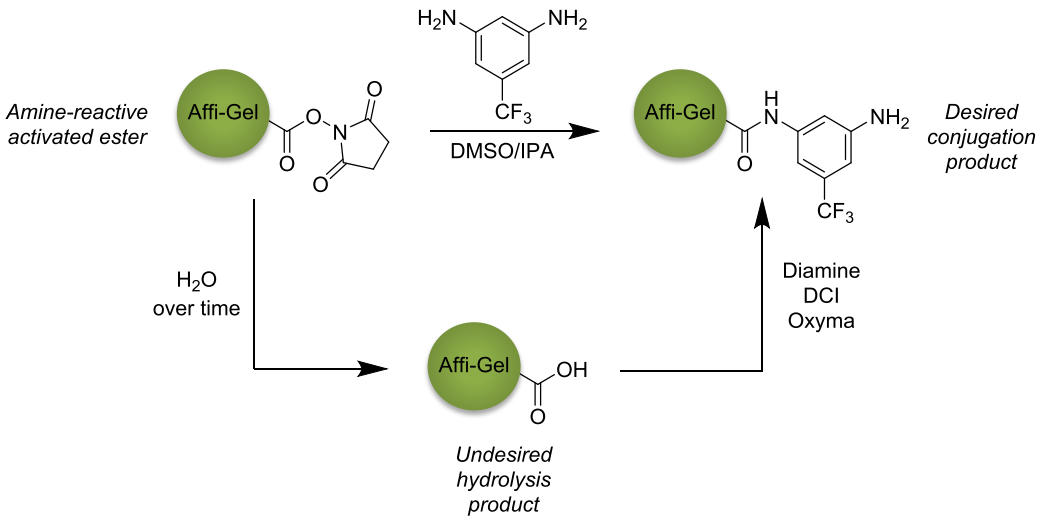
Figure 4-16 *Increasing the catch-and-release reaction time. (a) The yield of the catch-and-release reaction continued to improve until 48 h, suggesting the diazonium salt had not yet been quenched. (b) By ESI-MS, the 48 h product was found to be primarily doubly labelled. Shown is the MaxEnt deconvoluted spectrum, which will be discussed in the following section. The observed masses are shown with the increase in mass from the unmodified protein in brackets. Theoretical masses: 1 modification – 13,697 Da; 2 modifications – 13,712 Da; 3 modifications – 13,727 Da.*

Despite the improved yield with 48 h and 72 h reaction times, MS analysis of these protein samples revealed the presence of multiple modifications, whereas 24 h favoured single modification. Opting for selectivity over reactivity, the 24 h reaction time was adopted and this was justified by the high protein recovery allowed by the solid-phase system. Even in a low yielding reaction, there was a high recovery of unmodified protein in buffer with no additional reagents, allowing the user to recycle unreacted substrate.

4.3.4 Resin age

The yields for the optimised catch-and-release reaction were reproducible when using fresh resin but dropped off quickly upon storage of the resin after opening. The resin is functionalised with NHS-esters as described previously, and shipped for storage as a slurry in propan-2-ol. The manufacturers advise storage of the resin for up to one year in a -20 °C freezer in order to preserve 80% of original activity.[140] Each container had enough resin to perform ~25 catch-and-release experiments and once opened the containers were dated, sealed and stored in a -20 °C freezer. The most obvious route of degradation is by hydrolysis of the NHS-ester to give a carboxylic acid, which is not reactive to an amine at neutral pH without the addition of a coupling reagent.[160] Due to the difficulty in transferring the resin without drying it out, the propan-2-ol was replenished frequently during use and this could have introduced water to the container, increasing the rate of NHS-ester hydrolysis (Scheme 4-1). Resins of various ages (1 week; 1-2 months; >1 year) were used to assess the extent of degradation by performing the catch-and-release reaction simultaneously using each batch. Replicates of these experiments were performed with the coupling reagents DCI and Oxyma

used for the initial amide conjugation step, allowing coupling of the diamine to any carboxylic acid which had resulted from NHS-ester hydrolysis.



Scheme 4-1 *Recovery of aged resin by introduction of coupling reagents. Over time, the resin-bound NHS-ester is likely to undergo hydrolysis with water introduced from solvent replenishment. The resulting carboxylic acid is not reactive to the diamine alone, but the desired conjugation product can be achieved by activation of the carboxylic acid using DCI and Oxyma.*

In each reaction the diamine was added to the resin in DMSO:propan-2-ol, 10:90 (with or without the addition of coupling reagents) and agitated on a rotator for 4 h at rt. Using the standard protocol [no coupling reagents, Figure 4-17 (a): black bars], there was no significant difference in the performance of the 1 week and the 1-2 month old resins. However, these results represented ~70% reduction in yield when compared with fresh resin (246 µM). When using coupling reagents for diamine conjugation [Figure 4-17 (a): grey bars], the yield using the 1 week and 1-2 month old resins was recovered to >90% of that achieved with the fresh resin. These results suggest that after 1 week there was significant hydrolysis of the NHS-ester groups (perhaps exacerbated by replenishment of the propan-2-ol) which resulted in a drastic reduction in reactivity that was only recovered with the addition of peptide coupling reagents. The oldest resin used (date first opened was >1 year prior to reaction) gave a significantly reduced yield of modified protein. Even with the addition of coupling reagents, the yield was only recovered to 67% of the fresh resin. This suggests that the resin undergoes irreparable degradation beyond the 1 year shelf life as advertised, with the result that azobenzene formation was reduced (as observed by eye, the oldest resin was a less intense orange colour [Figure 4-17 (b)].

Despite the manufacturer's claim of resin stability in storage for up to 1 year, it was found that under current usage significant hydrolysis of the resin's NHS-esters occurred within one week of opening. To avoid this affecting future results, the whole resin volume was functionalised with 3,5-diaminobenzotrifluoride immediately upon opening in order to allow storage at -20 °C as the more stable amide.

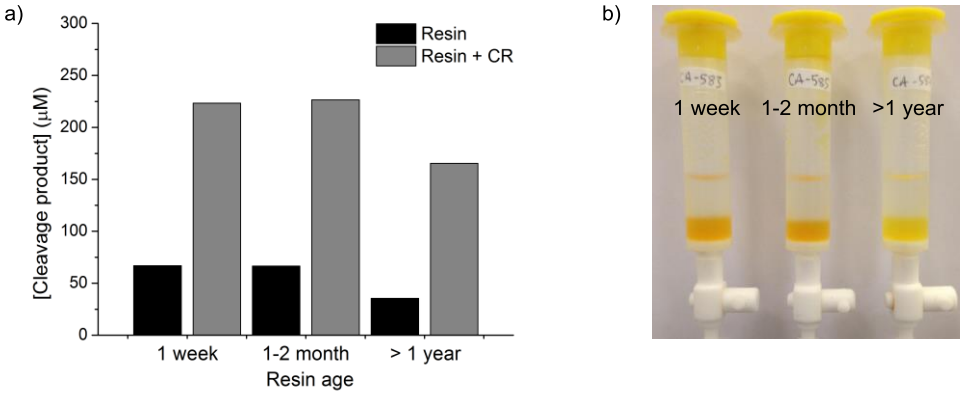


Figure 4-17 *NHS-ester hydrolysis in Affi-Gel 10. (a) The comparison of resins of three different ages, amide conjugation was performed in the absence (black) or presence (grey) of coupling reagents (CR). (b) The appearance of the three ages of resin upon azobenzene formation using the standard protocol (no coupling reagents) (left to right: 1 week; 1-2 months; >1 year).*

	[Cleavage product] (µM), % of fresh resin	
Resin age	No coupling reagent	With coupling reagents
Fresh	246, 100%	-
1 week	67, 27%	223, 91%
1-2 month	67, 27%	227, 92%
>1 year	36, 15%	166, 67%

Table 4-3 *The UHPLC-calculated concentrations of catch-and-release modified RNase A in three ages of resin in the absence and presence of coupling reagents.*

4.4 Catch-and-release of protein substrates

4.4.1 The relationship between SASA and yield

With an optimised protocol in hand, catch-and-release modification of a selection of proteins was undertaken. Given the solid-phase nature of the system, unbound protein was recovered in neutral pH buffer and in cases where the yield of modified protein was low, this sample could be recycled and reused if needed. Quantification of the unmodified and modified fractions by UHPLC and comparison with a calibration curve showed >93% total protein recovery in all cases [Figure 4-18 (a)].

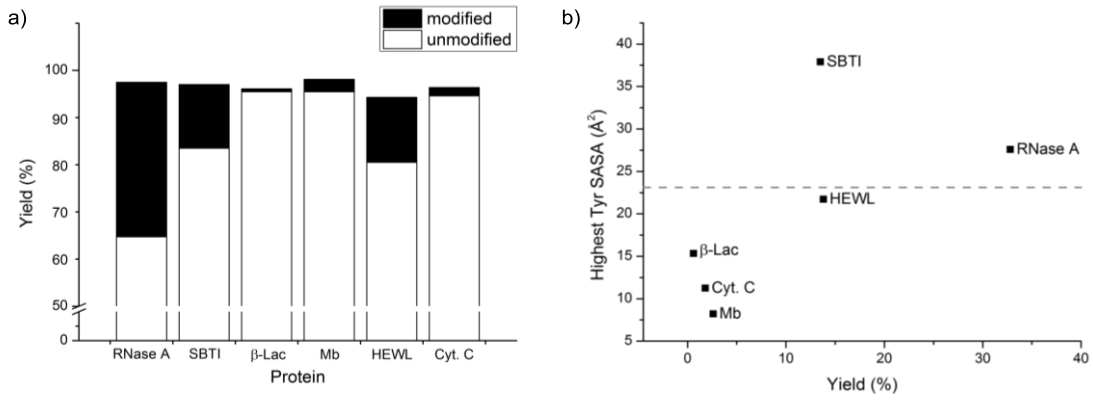


Figure 4-18 *Yield for the protein catch-and-release reactions. (a) While the yields for some of the substrates were low, accounting for recovered unbound protein resulted in >93% total protein recovery in all cases. (b) The yield % is plotted against the SASA of the most exposed Tyr residue in each protein, with the GYG benchmark given as a grey dashed line. The three highest yielding proteins each had a Tyr residue close to or above the SASA of the GYG benchmark.*

Protein	Yield (%)	Unbound protein (%)	Recovered yield (%)	Highest Tyr SASA (Å^2)
RNase A	32.8	64.7	92.9	27.6
SBTI	13.5	83.5	81.8	37.9
β-Lac	0.6	95.5	13.3	15.3
Mb	2.6	95.5	57.8	8.2
HEWL	13.8	80.5	70.8	21.7
Cyt. C	1.8	94.6	33.3	11.2

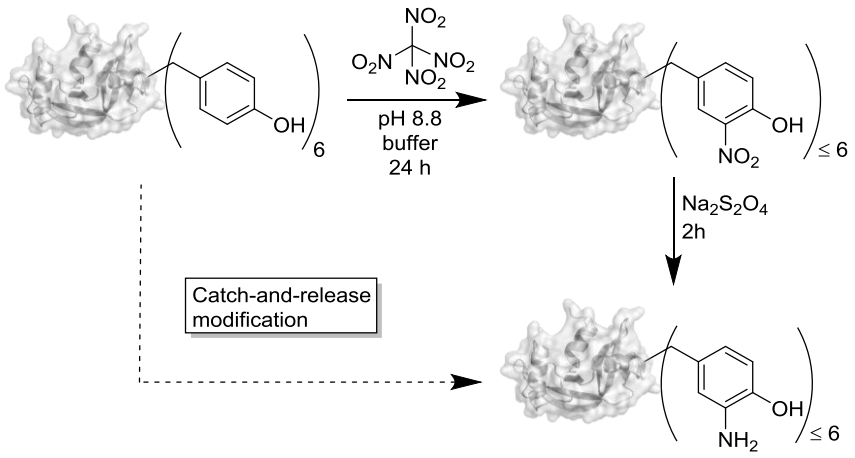
Table 4-4 *The modified and unmodified proteins were quantified using UHPLC and from this data, yields were calculated. The recovered yield is given as the yield of modified protein after subtraction of the recovered unmodified substrate. The SASA values of the most exposed Tyr residue in each substrate is also given.*

The yield of each reaction was plotted against the highest Tyr SASA for each protein [Figure 4-18 (b)]. Despite the limited scope of the protein screen, the three proteins which showed the

highest yields all had Tyr residues around or above the SASA of the GYG benchmark. RNase A was the highest yielding example and was therefore taken forward for MS analysis and oxidative coupling functionalisation.

4.4.2 An alternative route to *o*-aminophenol-modified RNase A

The *o*-aminophenol modification of Tyr has been previously explored in literature due to the role of Tyr nitration as a post-translational modification (PTM). This modification occurs by exposure of a Tyr residue to the radical nitrating agent nitrogen dioxide ($^{\cdot}NO_2$) which is produced in biological systems from nitrite- (NO_2^-) or peroxynitrite- ($ONOO^-$) dependant pathways.[161] In order to simulate this PTM, the small and highly reactive reagent tetranitromethane (TNM) has previously been used to modify RNase A.[147] This reagent has been shown to selectively nitrate the *ortho* position of the Tyr phenol ring to give 3-nitrotyrosine, mimicking the native PTM. Reduction of 3-nitrotyrosine to 3-aminotyrosine with $Na_2S_2O_4$ has been used as part of a modification strategy to quantify protein Tyr nitration, and this gives an alternative route to an *o*-aminophenol modified protein (Scheme 4-2).[122, 162] The TNM nitration followed by reduction strategy was selected as an interesting comparison to the catch-and-release methodology as the size and accessibility of the two Tyr-reactive species are completely juxtaposed. It can be assumed that TNM will have access to sites on the protein that the resin-tethered diazonium salt cannot reach, and thus we expect a different modification profile for the solution-phase and solid-phase approaches.



Scheme 4-2 *Modification of RNase A using the TNM nitration, reduction strategy.*

While it is sometimes described as a mild nitrating agent, TNM has also been investigated for use in explosives and as a fuel additive, leading one team of researchers who were investigating the mechanism of phenol nitration to comment:

"It should be noted that a mixture of TNM and pyridine has occasionally exploded several hours after mixing."[163]

With this in mind, TNM was handled with great care and reactions involving this reagent were performed behind a blast shield. A sample of RNase A was nitrated with 20 equivalents of TNM for 24 h. Upon treatment with $Na_2S_2O_4$, the 3-nitrotyrosine moiety was reduced to give 3-aminotyrosine, with a corresponding loss of the 357 nm absorbance peak (Figure 4-19). The resulting *o*-aminophenol modified RNase A was desalted by MWCO spin cartridge and assessed by MS for comparison with the catch-and-release modified sample.

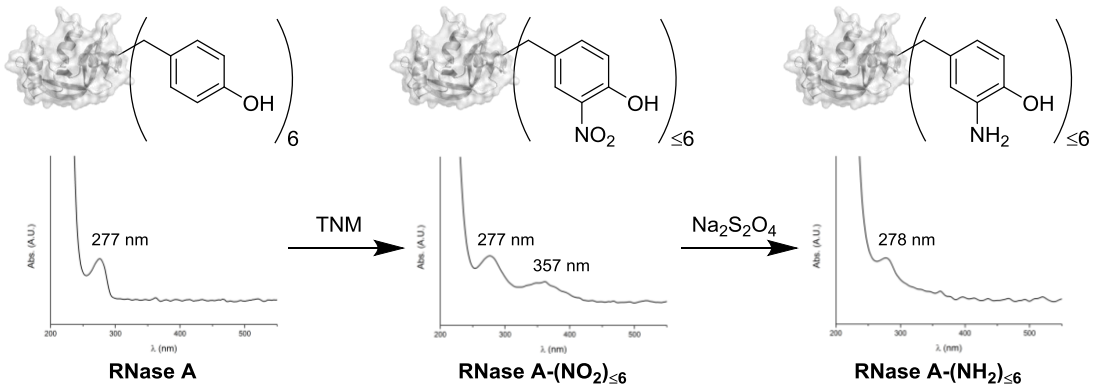


Figure 4-19 *The reduction of nitro-modified RNase A by $Na_2S_2O_4$ and the corresponding loss in absorbance at 357 nm.*

4.4.3 ESI-MS analysis of proteins

Compared to small molecules, the analysis of proteins by ESI-MS becomes complicated due to the increase in the number of ionisable sites. For this reason, it is common to observe high charge states which form a wide charge envelope (Figure 4-20). Using a deconvolution tool such as MaxEnt on MassLynx software, these multiple charge states can be deconvoluted to give the uncharged mass of the parent species (Figure 4-20). An unmodified RNase A standard was interrogated by this method, and the observed mass was found to be in excellent agreement with the theoretical mass based on the reported amino acid sequence.

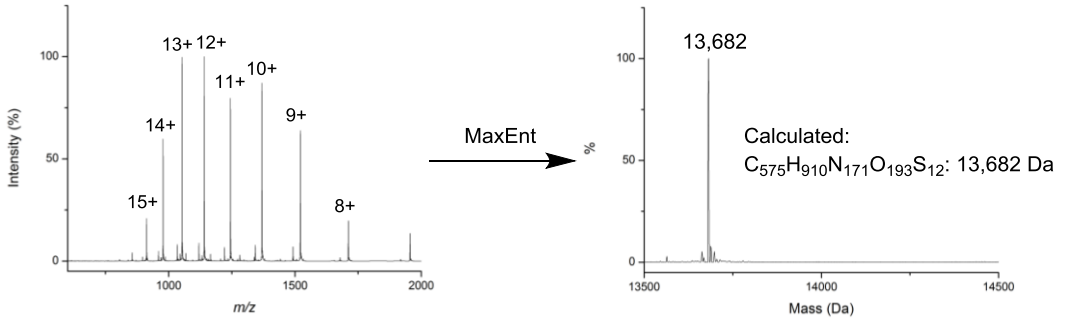


Figure 4-20 *ESI-MS of RNase A and the MaxEnt deconvolution tool. The broad charge envelope encompassing the 8+ to 15+ charge states is shown. This was deconvoluted using the MaxEnt tool in MassLynx software to give the uncharged parent mass, which agreed with the calculated mass of 13,682 Da.*

The unmodified RNase A standard was next compared with the solid-phase and solution-phase modified samples (Figure 4-21). The deconvoluted mass of each sample was calculated from the charge envelope, and all assigned peaks fell within 1 Da of the theoretical mass. The catch-and-release modified RNase A was found to be primarily singly modified, with the peak at 13,697 Da appearing most abundant. A small amount of unmodified and twice modified protein was also observed, however there were no significant peaks corresponding to higher levels of modification. This result supports the hypothesis that the extent of modification can be limited using a solid-phase approach, as we have achieved primarily single modification of a protein which contains six native targets. In contrast, the sample modified by solution-phase nitration and reduction showed higher levels of modification. The most abundant species at 13,741 Da corresponds to RNase A with four NH_2 modifications, and some of the 5 x modified species was also identified. There were no significant peaks suggesting modification of all six Tyr residues however, so it is possible that even the small nitrating reagent TNM failed to reach the most buried Tyr residue (Y097).

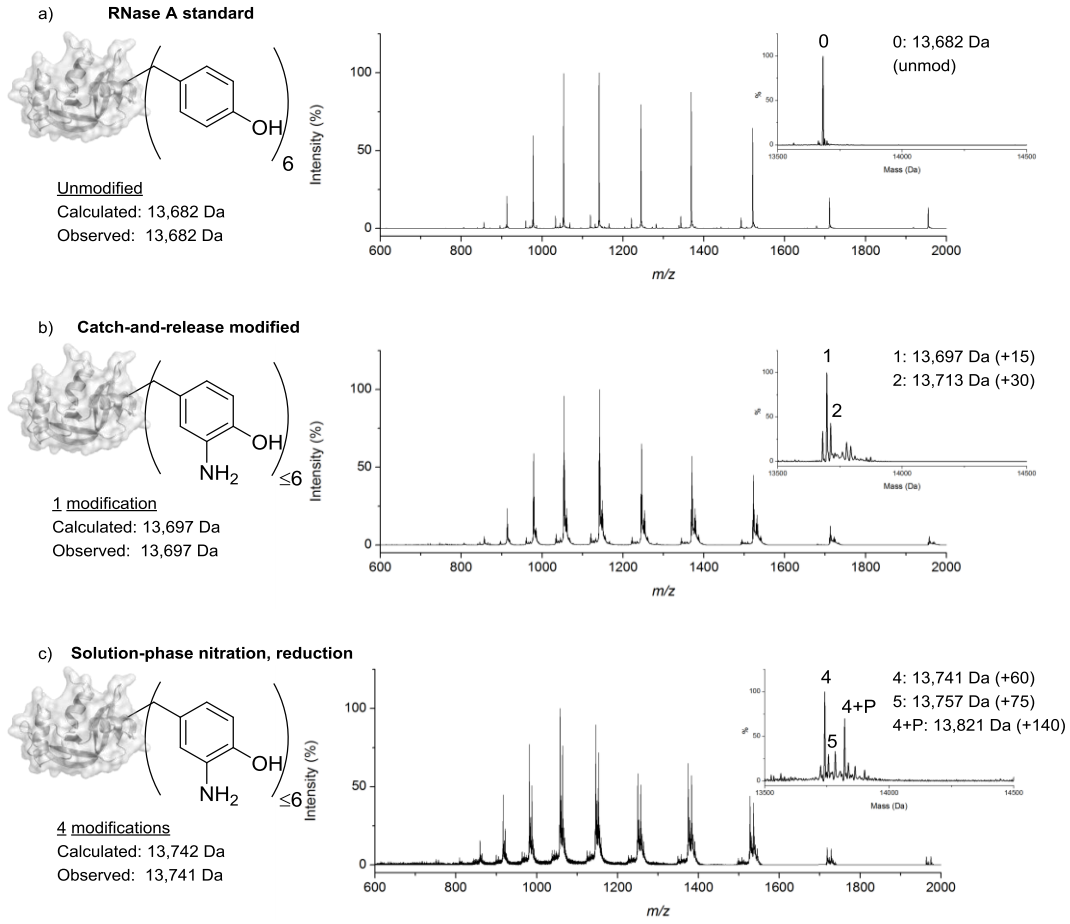


Figure 4-21 *Comparison of the solid-phase and solution-phase approaches to RNase A modification by ESI-MS. The charge envelope and deconvoluted mass (inset) are given for each sample. (a) The unmodified RNase A standard gave a deconvoluted mass of 13,682 Da. (b) The catch-and-release modified sample showed primarily the singly modified species at 13,697 Da. (c) The TNM-derived sample showed higher levels of modification, with the 4 x modified species appearing most abundant at 13,741 Da. The same peaks were also identified with phosphate adducts (+P).*

This work serves as a successful demonstration of the solid-phase approach and its ability to limit the extent of protein modification. In order to rationalise this in terms of accessible sites on the protein surface, it was necessary to identify the specific site, or sites, of modification in the protein sequence. This work was undertaken by top-down fragmentation of the denatured protein on the 12T FT-ICR instrument in the University of Edinburgh.[†]

[†] Top-down fragmentation of modified RNase A was performed by Dr C. Logan McKay, sample preparation and data analysis was performed by CA.

4.4.4 Top-down fragmentation of modified RNase A

In order to achieve good sequence coverage in top-down fragmentation techniques, the protein should be denatured to allow the exposure of buried residues. Unfortunately RNase A is a rather robust substrate which contains four disulphide bridges, and so these had to be reduced to allow sufficient unfolding of the protein. The reduction and denaturation of the protein was performed with TCEP and guanidine immediately prior to infusion into the MS. Initially fragmentation was attempted by collision-induced dissociation (CID), and the singly modified parent ion (*m/z* 1055.19, corresponding to the 13+ charge state) was isolated before being accelerated through a chamber containing neutral gas molecules. Upon collision, the energy transferred may cause fragmentation of the parent ion and these fragments are then analysed by MS. Using this technique, RNase A appeared only to be fragmented at the C-terminal end of the protein sequence and so fragmentation by electron-capture dissociation (ECD) was attempted instead. ECD is a relatively recent technique compared to CID, and while the exact mechanism of dissociation remains the subject of some debate, the event is initiated by protein capture of an electron resulting in formation of a radical species which is prone to bond cleavage.[164] ECD is a preferred method in certain circumstances as it is thought to be a fast and relatively mild approach to protein fragmentation. Interestingly, it is also known to cleave disulphide bonds which may explain the improved sequence coverage which resulted in this case. If TCEP reduction of the RNase A disulphides was incomplete, or if the resulting free thiols had re-oxidised before infusion, then this would hinder the fragmentation of the protein by CID, but shouldn't significantly affect fragmentation by ECD.

Following ECD fragmentation, the MS of the resulting peptides was recorded. Using the online tool Prosight PTM,[165, 166] peptide fragments were identified which matched the predicted *c*-/*z*- fragmentation pattern of *o*-aminophenol modified RNase A by ECD-MS. Using the software the +15 Da modification (from addition of the NH_2 group) was placed sequentially on each Tyr residue in the protein sequence. The sequence-coverage was assessed in each case by the number and placement of assigned peptides. By this method, the +15 Da modification was assigned to Y115, with sufficient coverage achieved to separate this position from the other Tyr residues in the protein (Figure 4-22).

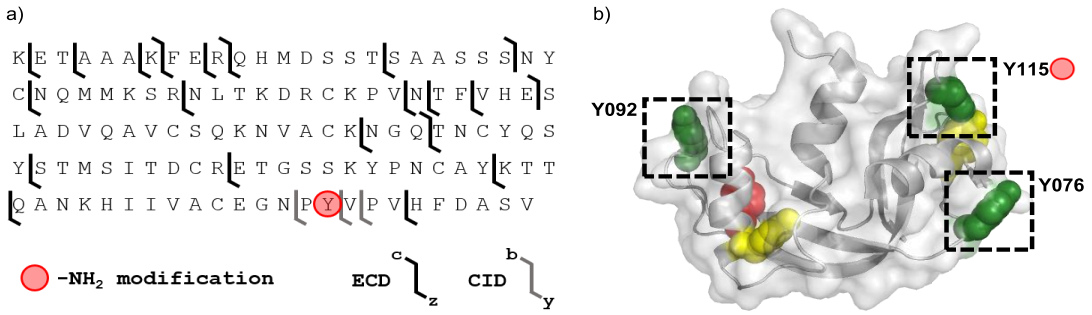


Figure 4-22 *Top-down fragmentation of catch-and-release modified RNase A. (a) The modified protein was fragmented by ECD and the modification was assigned to Y115, the third most exposed Tyr residue (red). (b) The three most exposed Tyr residues are shown in green. The relevant SASA values are Y076: 27.58 Å^2; Y092: 23.71 Å^2; Y115: 20.64 Å^2. (PDB ID: 4J5Z).*

The same process was applied to look for possible His modification, and it was observed that while there was no evidence for modification of three of the four native His residues, Y115 could not be distinguished from H105 by ECD fragmentation alone. Fortunately, despite the limited sequence coverage of CID fragmentation, this method did provide separation of the key C-terminal region, allowing distinction between H105 and Y115. Therefore the modification was assigned to Y115, and this result was also supported by bottom-up analysis.

Relating this back to Tyr surface-accessibility, it was interesting to note that Y115 was not the most exposed Tyr residue in the sequence. With a SASA value of 20.64 Å^2, the modified residue was still considered to be highly exposed (falling within close range to the GYG benchmark of 23.10 Å^2) however Y076 and Y092 had previously been identified as the two most exposed residues. In order to rationalise this, the local environment of each residue was explored to identify possible discrepancies between these three highly exposed positions.

4.4.5 Hydrogen bonding of Tyr residues

The first local effect which was considered to understand the difference in reactivity of the three Tyr residues Y076, Y092 and Y115 was H-bonding. Using the RNase A crystal structure (*PDB ID*: 4J5Z), these positions were examined to find polar contacts in which the Tyr hydroxyl group may take part.

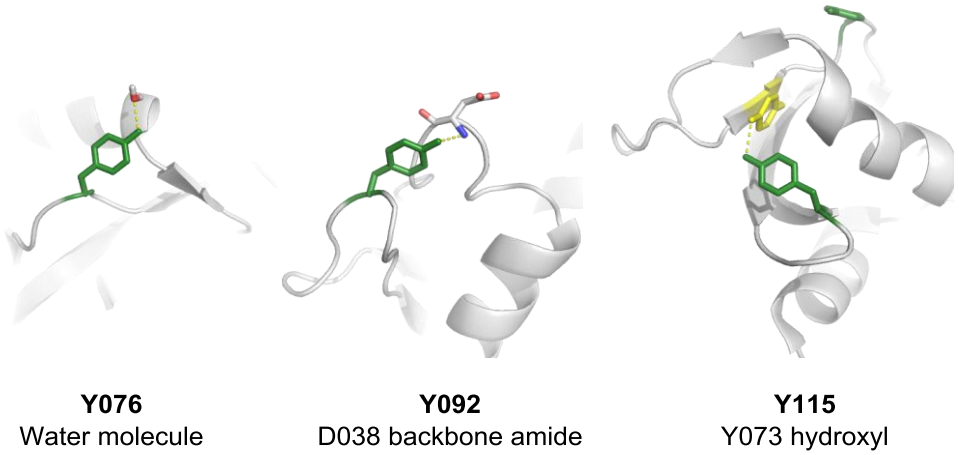


Figure 4-23 *Polar contacts of the three most exposed Tyr residues in RNase A. Y076 was found to have a polar contact with a water molecule in the crystal structure. Y092 made a polar contact with the backbone amide of a nearby aspartic acid residue. Y115 made a polar contact with the hydroxyl group of the intermediately exposed Y073 residue. (PDB ID: 4J5Z).*

The Y076 residue was found to make a polar contact with a nearby water molecule in the crystal structure, while the Y092 residue shared an interaction with the amide from the peptide backbone of an aspartic acid residue (D038). Y115 meanwhile was found to share an interaction with the phenolic OH of Y073. None of these interactions are of the type we would expect to enhance the reactivity of a Tyr residue; such an interaction would come for example from a carboxylate ion, which could hydrogen bond with the OH of the Tyr and reduce the pKa of the phenol (Figure 4-24). This would allow easier deprotonation of the Tyr to form the phenolate ion and drive nucleophilic attack from the *ortho* position of the ring in a general base catalysis-type mechanism.

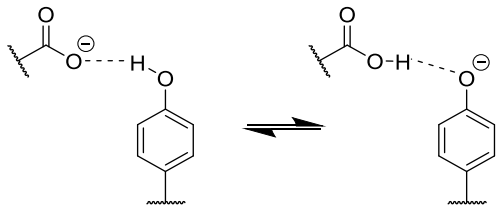


Figure 4-24 *Depiction of the type of intramolecular interaction likely to increase the reactivity of the Tyr residue by reducing the pKa of the side chain phenol.*

4.4.6 Electrostatics at the protein surface

Having assessed the polar contacts of the three most exposed Tyrs and found no clear difference in their potential reactivity, the electrostatic potential of the protein surface was next considered. Although the Affi-Gel resin bears a neutral linker to the NHS-ester group, once this has been functionalised as the amide-conjugated diazonium salt, a single resin bead will bear multiple positively-charged terminal groups. As such, it is possible that we would encounter some charge-charge repulsion between a positively charged face of the protein and the functionalised surface of the resin. Using PyMOL, the protein contact potential can be displayed using the vacuum electrostatics function and this can be used as a qualitative tool to explore the charged surface of a protein of interest. The function averages the electrostatic contributions of residues such as lysine, arginine, aspartic acid and glutamic acid which bear a charge at neutral pH, and plots this as a false-coloured surface on the protein.

The protein contact potential of RNase A was generated and the surfaces around the three most highly exposed Tyr residues (Y076, Y092 and Y115) were examined (Figure 4-25). Y076 was found to have a strong positively charged region located immediately next to the reactive *ortho* position of the phenol ring. In fact, this positive charge arises from K061, and it is the positively charged terminal amine of the K061 side chain which sits directly beside the reactive position of the phenol in the crystal structure. Despite the close proximity in space of Y092 and D038 (which bears a negatively charged side chain at neutral pH), Y092 forms a hydrogen bond to the backbone amide of this residue and the sidechain points away from the Tyr phenol ring. The negative potential of the D038 side chain also appears to be cancelled out by the positively charged Lys and Arg residues which sit on either side of it (K037 and R039). Furthermore, the Y092 residue has a Lys residue immediately next to it in the protein sequence (K091), making the surface surrounding this residue quite positively charged. Finally, there were no positively or negatively charged residues identified in close proximity to the Y115 residue, and in comparison to the two other highly exposed residues, this Tyr was found to sit on a surface which was relatively flat and uncharged. These findings support the hypothesis that charge-charge repulsion may play a role in the selectivity of the catch-and-release protein-tagging platform. Future work in this area could include a negatively charged functional group in the resin-bound construct to give an overall neutral resin once functionalised with a diazonium salt.

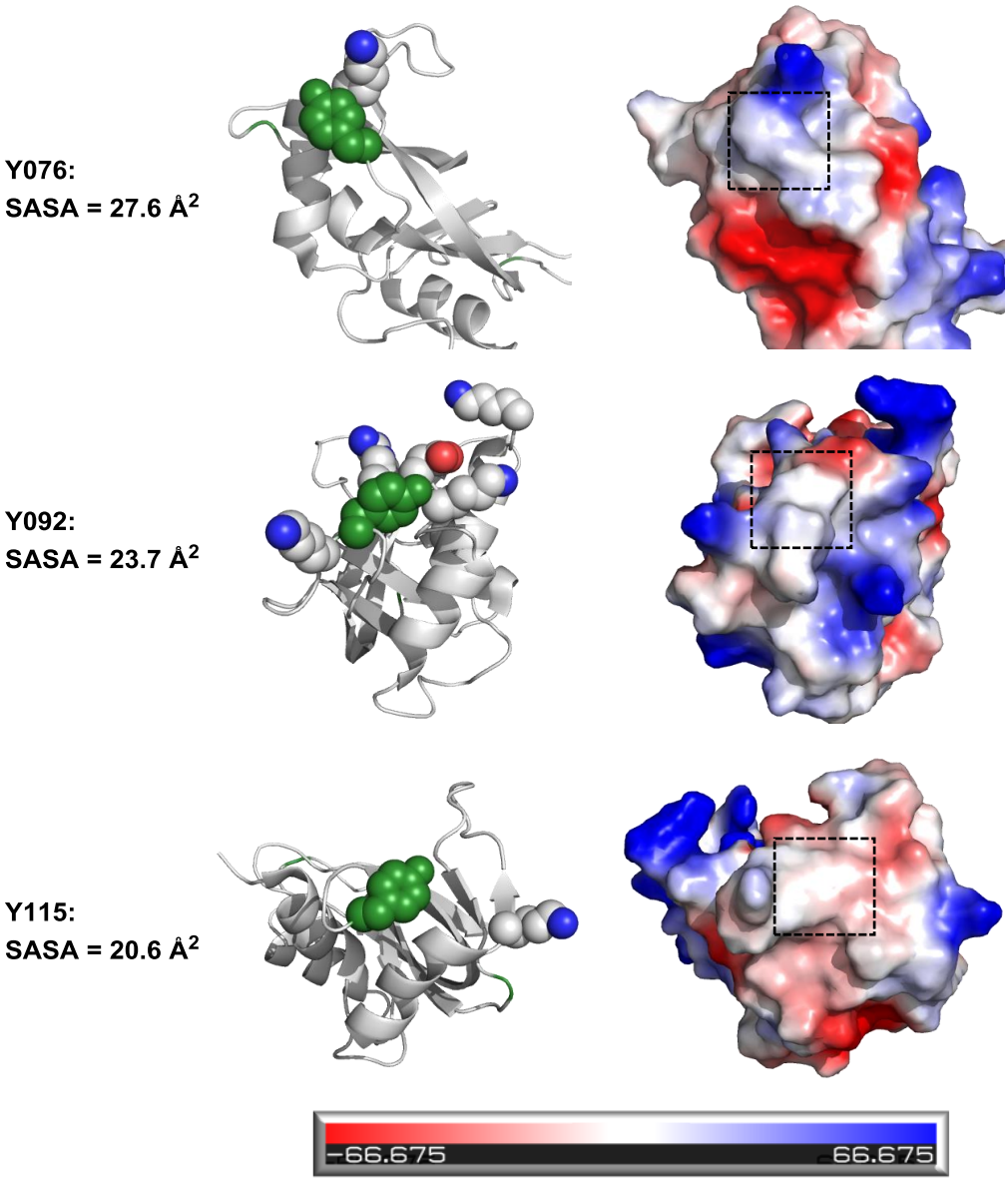


Figure 4-25 *Electrostatics of RNase A visualised on PyMol, Blue surfaces are more positively charged, red surfaces are more negatively charged. The most highly exposed Tyr residues (Y076 and Y092) were not found to be modified and this may be explained by their proximity to positively charged residues (Lys and Arg). In contrast, Y115 presents on a flat, mostly uncharged surface of the protein. (PDB ID: 4J5Z).*

4.5 Oxidative coupling on a protein substrate

In order to take advantage of the newly installed *o*-aminophenol functionality, the oxidative coupling strategy first reported by Francis *et al.*[122] was applied to the catch-and-release modified RNase A. As in the case of the catch-and-release chemistry, some adaptations were made to the procedure in moving from small molecules to proteins. In particular, a MWCO cartridge was used to separate the product from the small molecule reagents after completion of the reaction. The reagents themselves were also used in greater excess than in the small molecule example. Ten equivalents of the oxidising agent $K_3Fe(CN)_6$ had previously been used for oxidative coupling of small molecules, however this proved insufficient when applied to the protein substrate. This could perhaps be due to residual $Na_2S_2O_4$ which may not have been completely removed from the protein sample by desalting with a MWCO cartridge and, as a reducing agent, would potentially counteract the oxidising ferricyanide. The published conditions of 80 mM $K_3Fe(CN)_6$ were adopted for the protein samples,[122] representing over 1500-fold excess compared to the 50 µM protein. The anilines used for protein labelling were selected to have a UV absorbance which would allow detection of the product by HPLC-PDA even without a discernible change in R_t. These were used in a 20:1 stoichiometry to the protein giving a final concentration of 1 mM. The reaction was performed in pH 6.5 buffer for 18 h with agitation on a rotator at rt [Figure 4-26 (a)]. Following purification of the protein by successive washing and concentration on a MWCO cartridge, the protein products were identified by HPLC-PDA, bearing the distinctive absorbance peaks of the chosen fluorophores [Figure 4-26 (b)]. The fluorescein-conjugated sample was taken onto SDS-PAGE analysis to further confirm that the fluorophore was attached and not simply co-eluting with the protein peak on HPLC. By SDS-PAGE, an unmodified RNase A sample and a FITC-labelled RNase A sample (with up to ten fluorescent labels conjugated via Lys residues) were compared with the product of the oxidative coupling reaction [Figure 4-26 (c)]. The oxidative coupling product produced a band which was identified by Coomassie stain at the correct M_W for RNase A, and this band was also shown to be fluorescent, alongside the FITC-labelled standard. It should be noted that while a single fluorescein modification cannot be resolved from the unmodified sample by the position of the gel band (fluorescein modification M_W: 345 Da), the FITC-labelled sample contains up to ten fluorophores, resulting in a higher band on the gel and production of a strong signal when scanned for fluorescence. This work demonstrates the utility of an *o*-aminophenol modification as a synthetic handle, and together with the selective modification achieved through the catch-and-release reaction, this completes a novel protein-labelling strategy which targets Tyr residues in a selective manner.

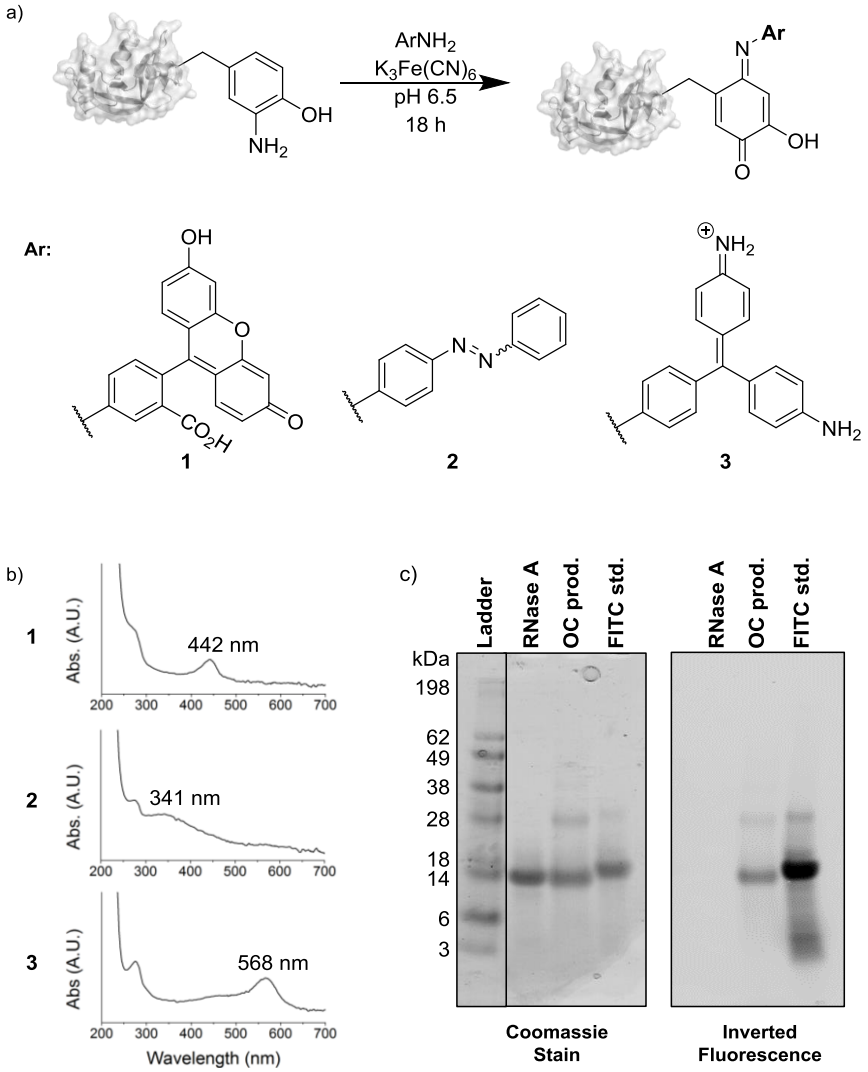


Figure 4-26 *Oxidative coupling (OC) of modified RNase A. (a) The reaction scheme for the oxidative coupling of an aniline to o-aminophenol-modified RNase A. Three fluorescent anilines ($ArNH_2$) were used: fluoresceinamine (1), aminoazobenzene (2) and pararosaniline (3). (b) PDA analysis of the OC products gave new absorbance peaks resulting from incorporation of the fluorophore (1: 442 nm, 2: 341 nm, 3: 568 nm). (c) SDS-PAGE of the fluoresceinamine OC product with coomassie stain and inverted fluorescence. The OC product was compared with unmodified RNase A and a FITC-labelled standard.*

4.6 Conclusion

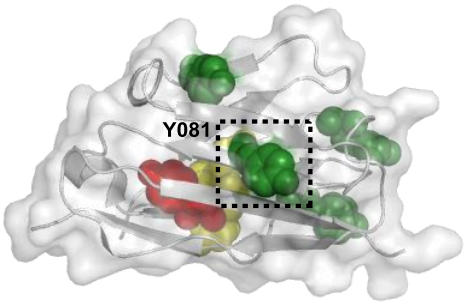
To conclude, this Chapter has built on the small molecule and peptide work shown in Chapters 2 and 3, both in solution and on solid-phase, to deliver a protocol for the selective modification of a native sequence protein which contains multiple Tyr residues. Using *in-silico* tools, it has been possible to assess the availability of each potentially reactive residue through the application of the SASA calculation. By the development of an automated script to perform this duty, and through the expansion of the code to encompass other potentially targetable residues, this method may be used in future protein-tagging applications. Following catch-and-release of multiple protein substrates, it was observed that the highest yielding examples displayed a highest Tyr SASA close to or above the GYG benchmark, and it would be interesting in future work to explore this relationship through expansion of the number of protein candidates and in-depth MS analysis of the modified product.

In RNase A it was found that the most exposed Tyr was not in fact the residue which had been labelled, instead labelling of another highly exposed Tyr was achieved. Through consideration of intramolecular interactions and protein surface electrostatics, there is an avenue to refine future predictions of reactivity, especially if these functions can be incorporated into the automated SASA scripts.

Oxidative coupling of a fluorescent aniline to the catch-and-release modified RNase allowed visualisation of the labelled product by PDA and in-gel fluorescence scanning. In future, the nature of the aniline may be changed to achieve protein labelling with a PEG chain, a biotin group or another functional small molecule, increasing the application of the current work.

4.7 Future work

Future work will primarily involve application of the developed catch-and-release system to research-appropriate substrates. As a potential target, the *anti*-GFP nanobody is a substrate of interest, as labelling this with a fluorophore would allow co-localisation microscopy experiments with GFP as a proof of principle. SASA analysis of α-GFP from the published co-crystal structure with GFP (Figure 4-27) suggests that modification of the most exposed Tyr residue, Y081 would not interfere with binding, as it is located on the opposite face of the protein from the binding interface.



Residue	Avg. SASA (Å^2)	Std. Dev.
Y033	17.02	1.55
Y038	15.52	0.78
Y061	17.03	0.63
Y081	21.74	1.65
Y095	0.00	0.00
Y096	6.89	0.49
Y105	17.15	2.94

Figure 4-27 *SASA evaluation of the 7 Tyr residues in the anti-GFP nanobody (PDB ID: 3OGO). The SASA value of each Tyr residue was averaged over the four chains contained in the crystal structure, the results ranged from the buried Y095 residue to the most exposed Y081, which sits on the opposite face of the protein from the GFP binding interface.*

Beyond α-GFP, nanobodies are a class of substrates which would be generally suitable for fluorescent labelling, as they would provide a small fluorescent probe which is preferred for modern super-resolution imaging techniques. Unfortunately, few crystal structure deposits of nanobodies currently exist, limiting the use of the SASA calculation in this instance.

Alternatively, the catch-and-release system could be used for the reversible immobilisation of proteins for purposes such as biophysical assays. By immobilisation of a protein substrate on an SPR chip, this could prove a useful means of characterising protein interactions, while

selective modification would provide a homogeneous display of the immobilised protein as each biomolecule would be anchored to the support by the most reactive Tyr residue.

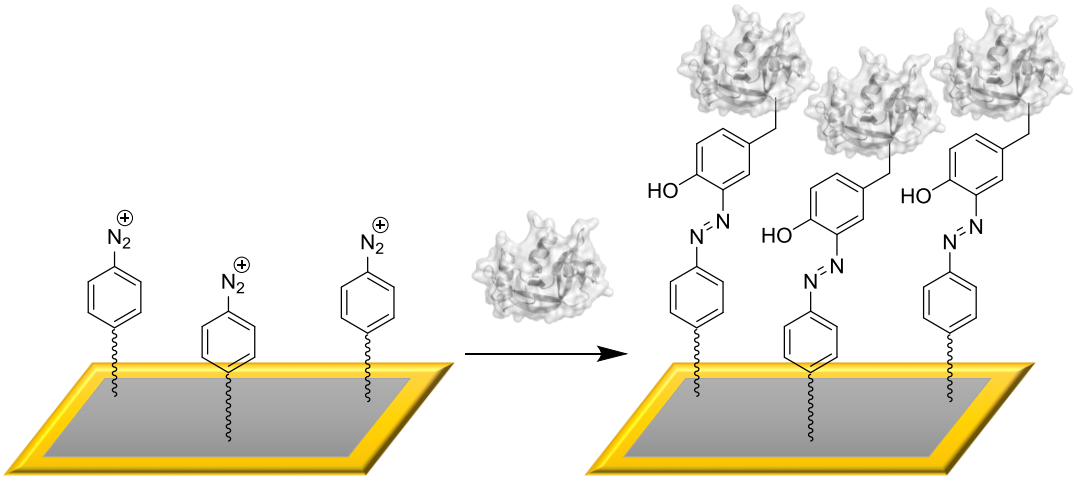


Figure 4-28 *A potential application of the catch-and-release chemistry is in the reversible immobilisation of proteins for use in biophysical assays.*

Finally, the solid-phase modification platform and cleavable linker design could be used to expand the application of catch-and-release modification to other amino acids. Using a Tyr-inspired three-vector linker, a range of reactive groups could be used for protein capture, before cleavage with dithionite and functionalisation through a click handle.

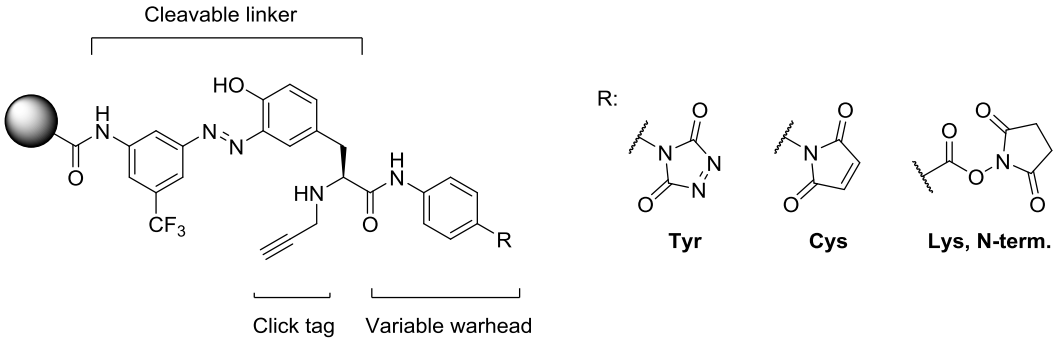


Figure 4-29 *A possible expansion of the catch-and-release approach to encompass other amino acid targets. This approach would retain the linker design established in this work, with the addition of a conventional click handle such as an alkyne, and a variable site to incorporate a range of other amino acid targeting functional groups.*

Chapter 5 Experimental

5.1 General Methods

All non-aqueous reactions were performed under nitrogen atmosphere using oven-dried glassware that was cooled in a desiccator prior to use. MeOH, DCM and acetonitrile were dried and purified by a solvent purification system using activated alumina. Unless otherwise stated, all reagents were commercially sourced and used without further purification. Organic layers were dried using magnesium sulphate unless otherwise stated. Saturated aqueous solutions of inorganic salts are represented as (volume, sat. aq.). Triethylamine was distilled over CaH_2 prior to use.

Thin layer chromatography was performed using Merck silica gel 60F254 foil-backed plates with visualisation under ultraviolet light (254 nm), or using $KMnO_4$ (aq.) or ninhydrin staining. R_f values are given, with the solvent system quoted as a v:v ratio. Column chromatography was performed using Merck Kieselgel 60 (Merck 9385) silica under positive pressure from a hand pump. Eluent compositions are given as v:v ratios.

Melting points were determined on a Gallenkamp Electrothermal melting point apparatus and are uncorrected.

Optical rotations were measured on a PolAAr 20 series polarimeter with a path length of 1.0 dm at the sodium D line (589 nm) and are reported as follows: $[\alpha]_D$, concentration (g/100 mL), solvent. All optical rotations were measured at room temperature.

Infra-red spectra were recorded on a Shimadzu IRAffinity-1 machine. The samples were analysed neat, and wavelengths of maximum absorbance (υ_{max}) are quoted in cm^{-1}.

^{1}H and ^{13}C NMR spectra were obtained on Bruker instruments at the stated frequency. The data is presented as follows: Chemical shift (in ppm on the δ scale relative to δTMS = 0 ppm), integration, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, quin = quintet, m = multiplet, td = triplet of doublets, dd doublet of doublets, dt = doublet of triplets, br = broad), coupling constants (in Hertz, Hz) and the assigned peaks are shown in *italics* (e.g. CO_2*H*).

Small molecule and peptide products were analysed by ESI-MS using a Bruker MicroTOF2 instrument. Mass-to-charge ratios (*m/z*) of parent peaks and significant fragments are identified, along with their relative intensities (%).

Protein analysis by SDS-PAGE was performed using a 12 well 4-12% NuPAGE Bis-Tris Gel (Novex) in a Bolt Mini Gel Tank (Novex). The running buffer was NuPAGE MES SDS, the M_W ladder used was SeeBlue 4-250 kDa (Invitrogen). The protein samples were mixed with Laemmli buffer and heated to 100 °C for 5 min, before loading into the sample wells. The gel was developed for 35 min at 200 V. Fluorescent bands were detected using a Typhoon 9400 Variable Mode Imager (Amersham Biosciences). The same gel was then stained with InstantBlue (Expedeon) and developed for 15 min on a shaking table.

5.2 LC Methods

5.2.1 U/HPLC-PDA methods

Reaction monitoring and analysis was performed by reverse-phase chromatography using either a HPLC with photo-diode array detector or a UHPLC with photo-diode array detector. HPLC was performed using a Waters 600E pump, a Waters 717plus autosampler and a Waters 996 PDA equipped with a Phenomenex SphereClone ODS C18, 5 µm, 4.6 x 150 mm column (System 1). UHPLC was performed on a Waters Acquity UPLC PDA equipped with a Waters Acquity UPLC BEH300 C4, 1.7 µm, 2.1 x 100 mm column (System 2) or a Waters Acquity UPLC BEH130 C18, 1.7 µm, 2.1 x 100 mm column (System 3). The U/HPLC traces were exported as ASCII data using Empower software by Waters. The traces were processed using Origin (version 8.5.1) by OriginLab Corporation, and baseline subtraction was performed using the trace obtained from a blank injection. Retention time (R_t) was taken from the unprocessed spectrum; where conversion by U/HPLC is given, this was calculated using the peak area calculation tool in Empower on the unprocessed spectrum at 210 nm absorbance.

Method 1: System 1; flow rate 1 mL min^{-1}; ambient column temperature, total method time 20 min; PDA absorbance spectrum recorded from 200-800 nm.

Time (min)	**% H_2O (+0.1 % v/v TFA)**	**% MeCN (+0.1 % v/v TFA)**
0	95	5
10	5	95
12	5	95
15	95	5

Method 2: System 2 or System 3; flow rate 0.2 mL min^{-1}; column temperature 60 °C; total method time 12 min, PDA absorbance spectrum recorded from 200-400 nm.

Time (min)	% H_2O (+0.1 % v/v TFA)	% MeCN
0	95	5
0.5	95	5
10.5	5	95
10.9	5	95
11	95	5

5.2.2 Preparative HPLC method

Preparative RP-HPLC was performed using a Waters 600E pump with a Waters 486 tuneable absorbance detector, using a Phenomenex Luna C18(2), 5 µm, 21.2 x 250 mm column (System 4). Isocratic solvent systems were used for purification using H2O + 0.1% TFA and MeCN + 0.1% TFA; the flow rate was 18 mL min^{-1}. Samples were dissolved in DMSO, and ≤ 0.25 mL was injected by Rheodyne.

5.2.3 UHPLC-MS method

MS analysis of intact protein samples was performed using a Waters Acquity UPLC equipped with a Waters Acquity UPLC BEH300 C4, 1.7 µm, 2.1 x 100 mm column and a Waters SYNAPT G2 mass spectrometer (System 5). The UHPLC-MS method was modelled on the UHPLC-PDA method, with an initial 2.5 min of equilibration at which time the UHPLC outlet was diverted to waste.

Method 3: System 5; flow rate 0.2 mL min^{-1}; column temperature 60 °C; total method time 15 min.

Time (min)	% H_2O (+0.1 % v/v FA)	% MeCN (+0.1% v/v FA)
0	95	5
2.5	95	5
12.5	5	95
12.9	5	95
13	95	5

5.3 Solid-phase peptide synthesis

Resin preparation

All peptides were synthesised using Rink Amide AM resin (200-400 mesh, 0.69 mmol/g). The resin was weighed into a SPE tube, and swelled in DCM for 30 min with agitation on a rotator at rt. The resin was drained by vacuum filtration and washed with DMF. The Fmoc-protected resin was deprotected using 20% piperidine in DMF with agitation on a rotator for 10 min at rt.

TNBS free amine test

After each coupling and deprotection step, a small portion of resin beads was removed and added to an Eppendorf with a small volume of DMF. Using a plastic pastette, ~5 drops of DIPEA solution (10% in DMF) and 1 drop of TNBS solution (5% aq.) were added. The Eppendorf was agitated on a rotator for 10 min at rt and subjected to centrifugation. The beads were washed with DMF and visual inspection determined the presence of free amine (red beads) or the absence of free amine (colourless beads).

Coupling of amino acids

On-resin amide coupling was performed by mixture of an Fmoc-protected amino acid (3 equiv.), Oxyma (3 equiv.) and DIC (3 equiv.) in DMF, which was then added to the resin and agitated on a rotator for 1 h at rt. The resin was drained by vacuum filtration and excess reagents were washed from the solid phase (3 x DMF, 3 x DCM, 1 x DMF). The TNBS test

was performed on a small portion of resin beads, and the coupling step was repeated if any red beads were observed.

Fmoc deprotection

Fmoc deprotection was performed by addition of piperidine (20% in DMF) to the resin, followed by agitation on a rotator for 10 min at rt. The resin was drained and washed with DMF and the TNBS test was performed on a small portion of resin beads. The deprotection step was repeated if any colourless beads were observed.

Acetyl capping

Following completion of the desired peptide sequence, the N-terminus was protected as an acetamide by addition of acetic anhydride (20 equiv.) and DIPEA (20 equiv.) to the resin, with agitation on a rotator for 1 h at rt. The resin was washed (3 x DMF, 3 x DCM, 1 x DMF) and the TNBS test was performed on a small portion of resin beads, the capping step was repeated if any red beads were observed.

Cleavage and global deprotection

The synthesised peptide was cleaved from the resin by addition of the cleavage cocktail (TFA:TIS:H_2O, 95:2.5:2.5 by volume) and agitation on a rotator for 3 h at rt. Peptides containing a Cys residue were cleaved with a cocktail containing a disulphide-reducing agent (TFA:TIS:H_2O:EDT, 92.5:2.5:2.5:2.5 by volume). The eluate from the resin was collected in a Falcon tube and the resin was washed with TFA. TFA was removed under a stream of N_2, and the peptide was triturated with cold diethyl ether. The peptide was purified by preparative RP-HPLC, and lyophilised to give the final product as a colourless amorphous solid.

Resin storage

If overnight storage was required before completion of the synthesis, the immobilised Fmoc-protected peptide was washed (4 x MeOH) and suspended in MeOH for 10 min. The resin was drained and washed with diethyl ether (3 x), before drying by vacuum filtration and storage at -20 °C. The resin was swelled as before prior to continuation of peptide synthesis.

5.4 SAv catch-and-release methods

5.4.1 Small molecule and peptide substrates

SAv resin (0.2 mL settled resin, ≥ 17 nmol biotin binding capacity) was added to a 3 mL SPE tube fitted with a polyethylene frit and a tap. The resin was equilibrated with 2 CV buffer (phosphate buffer, 100 mM, pH 7.2) and the biotinylated aniline **33** (0.3 mL, 10 mM in DMSO:buffer, 10:90, 3 µmol) was added. The resin was agitated on a rotator at rt for 2 h and excess aniline was washed from the resin using buffer (~10 CV, the eluate was analysed by UHPLC until the aniline peak had disappeared). $NaNO_2$ (21 mg, 304 µmol) was mixed with acetic acid solution (0.3 mL, 0.6 M) and added to the resin. The reaction mixture was agitated for 1 h at rt, and the resin was then washed with buffer (2 CV), before addition of the Tyr-bearing substrate (0.3 mL, 1 mM, 300 nmol). The reaction mixture was agitated on a rotator for 18 h at rt. The resin was then washed with buffer (20 CV, the eluate was analysed by UHPLC until the substrate peak had disappeared). Freshly prepared $Na_2S_2O_4$* (0.6 mL, 100 mM. 60 µmol) was added to the resin and the mixture was agitated at rt for 18 h. The resin was drained and washed with buffer (5 CV) to release the modified substrate. The product was desalted using a C_{18} flash cartridge, with elution in MeOH.

5.4.2 Myoglobin catch-and-release method

SAv resin (0.2 mL settled resin, ≥ 17 nmol biotin binding capacity) was added to a 3 mL SPE tube fitted with a polyethylene frit and a tap. The resin was equilibrated with 2 CV buffer (phosphate buffer, 100 mM, pH 7.2) and the biotinylated aniline **33** (0.3 mL, 10 mM in DMSO:buffer, 10:90, 3 µmol) was added. The resin was agitated on a rotator at rt for 2 h and excess aniline was washed from the resin using buffer (~10 CV, the eluate was analysed by UHPLC until the aniline peak had disappeared). $NaNO_2$ (21 mg, 304 µmol) was mixed with acetic acid solution (0.3 mL, 0.6 M) and added to the resin. The reaction mixture was agitated for 1 h at rt, and the resin was then washed with buffer (2 CV), before addition of Mb (0.3 mL, 0.6 mM, 180 nmol). The reaction mixture was agitated on a rotator for 18 h at rt. The resin was then washed with buffer (20 CV, the eluate was analysed by UHPLC until the substrate peak had disappeared). Freshly prepared $Na_2S_2O_4$ (0.6 mL, 100 mM. 60 µmol) was added to the resin and the mixture was agitated at rt for 18 h. The resin was drained and washed with buffer (5 CV) to release the modified substrate. The product was desalted using a size-exclusion spin cartridge.

* $Na_2S_2O_4$ has been reported to deteriorate in the presence of atmospheric H_2O or in aqueous solution, [122] it was therefore stored in a desiccator and solutions were prepared immediately prior to use.

5.5 Affi-Gel catch-and-release methods

5.5.1 Small molecule and peptide substrates

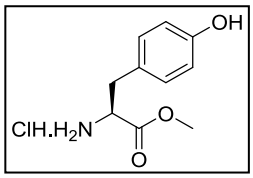
Affi-gel 10 resin (1 mL settled resin in IPA) was added to a 5 mL SPE tube fitted with a polyethylene frit and a tap. *p*-Phenylenediamine (17 mg, 157 µmol) was dissolved in IPA (3 mL) and added to the resin, which was then agitated on a rotator for 4 h at rt. The resin was washed with MeOH (5 CV), before addition of ethanolamine (10 µL, 166 µmol) in IPA (3 mL) to block any remaining NHS-esters. The resin was agitated for 1 h at rt and washed with IPA (3 CV) and water (3 CV). $NaNO_2$ (35 mg, 507 µmol) was mixed with acetic acid solution (3 mL, 0.6 M) and added to the resin. The reaction mixture was agitated for 1 h at rt, and then washed with buffer (2 CV phosphate buffer, 100 mM, pH 7.2), before addition of the Tyr-bearing substrate (3 mL, 5 mM in buffer, 15 µmol). The reaction mixture was agitated for 18 h at rt. The resin was drained and washed with buffer (10 CV, eluate was analysed by UHPLC until the substrate peak had disappeared). $Na_2S_2O_4$ (3 mL, 300 mM) was added and the resin was mixed on a rotator for 18 h. The modified substrate was collected by draining and washing the resin (2 CV H_2O), the product was desalted using a C_{18} flash cartridge, with elution in MeOH.

5.5.2 Protein substrates

Affi-gel 10 resin (1 mL settled resin in IPA) was added to a 5 mL SPE tube fitted with a polyethylene frit and a tap. An aromatic diamine (*p*-phenylenediamine, 3,5-diaminobenzoic acid or 5-(trifluoromethyl)-1,3-phenylenediamine, 150 µmol) was dissolved in DMSO:IPA (3 mL, 10:90) and added to the resin, which was then agitated on a rotator for 4 h at rt. The resin was washed with MeOH (5 CV), before addition of ethanolamine (10 µL, 166 µmol) in IPA (3 mL) to block any remaining NHS-esters. The resin was agitated for 1 h at rt and washed with IPA (3 CV) and water (3 CV). $NaNO_2$ (35 mg, 507 µmol) was mixed with acetic acid solution (3 mL, 0.6 M) and added to the resin. The reaction mixture was agitated for 1 h at rt, and then washed with buffer (2 CV phosphate buffer, 100 mM, pH 7.2), before addition of the protein substrate (3 mL, 1.95 mM in buffer, 5.85 µmol). The reaction mixture was agitated for 18 h at rt. The resin was drained and washed with buffer (25 CV, eluate was analysed by UHPLC until the substrate peak had disappeared). $Na_2S_2O_4$ (3 mL, 300 mM) was added and the resin was mixed on a rotator for 18 h. The modified substrate was collected by draining and washing the resin (1 CV buffer), the product was desalted using a MWCO spin cartridge.

5.6 Experimental procedures for Chapter 2

Methyl (2*S*)-2-amino-3-(4-hydroxyphenyl)propanoate hydrochloride, **1**

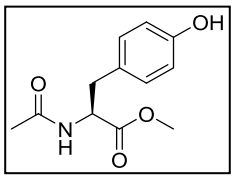


MeOH (14 mL) was cooled to 0 °C in an ice bath and acetyl chloride (0.80 mL, 11.3 mmol) was added dropwise (gas evolved). The solution was stirred at 0 °C for 15 min, before portion-wise addition of L-tyrosine (0.68 g, 3.77 mmol). The reaction mixture was allowed to warm to room temperature, then heated to reflux for 3 h. Solvent was removed *in vacuo* and the resultant solid was recrystallised from MeOH to give the methyl ester **1** as colourless needles (0.64 g, 74%).

R_f (DCM:MeOH, 80:20) = 0.66; **mp** 185-187 °C, MeOH, lit.[167] 188-189 °C, MeOH; **$[\alpha]_D$ =** -6.0 (c 0.50, H_2O) lit.[167] -5.2 (c 2.40, H_2O); **IR** (neat, cm^{-1}) 3335 (OH), 1742 (C=O) 1614 (C=C), 1591 (C=C), 1514 (C=C); **^{1}H NMR** δ (500 MHz, CD_3OD) 7.09 (2H, d, *J* = 8.5 Hz, Ar*H*), 6.81 (2H, d, *J* = 8.5 Hz, Ar*H*), 4.26 (1H, dd, *J* = 7.5, 5.9 Hz, NH_2CH), 3.84 (3H, s, CO_2CH_3), 3.19 (1H, dd, *J* = 14.5, 5.9 Hz, CH_AH_B), 3.09 (1H, dd, *J* = 14.5, 7.5 Hz, CH_AH_B); **^{13}C NMR** δ (126 MHz, CD_3OD) 169.2 (*C*O), 157.1 (Ar*C*), 130.1 (2Ar*C*H), 124.2 (Ar*C*), 115.5 (2Ar*C*H), 54.0 (*C*H), 52.2 (CH_3), 35.3 (CH_2); ***m/z*** (ESI+, MeCN/H_2O) 218 ($[M+Na]^+$, 30%), 196 ($[M+H]^+$, 100), 179 ($[M-OH+H]^+$, 80), 136 ($[M-CO_2CH_3]^+$, 50).

^{1}H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[167]

Methyl (2*S*)-2-acetamido-3-(4-hydroxyphenyl)propanoate, **2**



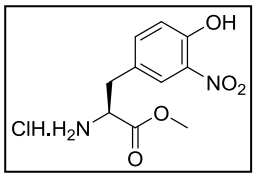
The ammonium salt **1** (1.75 g, 7.56 mmol) was dissolved in water (10 mL) and cooled to 0 °C in an ice bath. NaOAc (15.1 mL, 75.5 mmol; 5 M aq.) was added to the reaction mixture, followed by acetic anhydride (1.07 mL, 11.3 mmol). The reaction was monitored by TLC until judged to have reached completion (30 min) by which time a white precipitate had formed. The reaction mixture was filtered, the precipitate was washed with diethyl ether (2 × 20 mL) and dried *in vacuo* to give the acetamide **2** as a colourless solid (1.73 g, 97%).

R_f (DCM:MeOH, 80:20) = 0.76; **mp** 130-132 °C, lit.[168] 136-137 °C; **$[\alpha]_D$** = +32.0 (c 1.00, MeOH) lit.[168] +29.7 (c 0.41, MeOH); **IR** (neat, cm^{-1}) 3547 (NH), 3319 (OH), 1728 (C=O), 1657 (C=O), 1614 (C=C), 1545 (C=C), 1514 (C=C); **^{1}H NMR** δ (500 MHz, CD_3OD) 7.03 (2H, d, *J* = 8.5 Hz, 2Ar*H*), 6.72 (2H, d, *J* = 8.5 Hz, 2Ar*H*), 4.60 (1H, dd, *J* = 8.6, 5.9 Hz, NHC*H*), 3.70 (3H, s, CO_2CH_3), 3.05 (1H, dd, *J* = 13.9, 5.9, CH_AH_B), 2.87 (1H, dd, *J* = 13.9,

8.6 Hz, CH_AH_B), 1.93 (3H, s, CH_3CO); **^{13}C NMR** δ (126 MHz, CD_3OD) 172.3 (*C*O), 171.8 (*C*O), 156.0 (Ar*C*), 129.8 (2Ar*C*H), 127.4 (Ar*C*), 114.8 (2Ar*C*H), 54.2 (*C*H), 51.2 (OCH_3), 36.3 (CH_2), 20.8 (CH_3); ***m/z*** (ESI+, MeCN/H_2O) 260 ($[M+Na]^+$, 100%), 238 ($[M+H]^+$, 5).

^{1}H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[168]

Methyl (2*S*)-2-amino-3-(4-hydroxy-3-nitrophenyl)propanoate hydrochloride, **3**

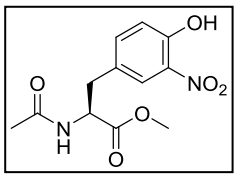


MeOH (7 mL) was cooled to 0 °C in an ice bath and acetyl chloride (0.190 mL, 2.67 mmol) was added dropwise (gas evolved). The solution was stirred at 0 °C for 15 min, before portion-wise addition of 3-nitro-L-tyrosine (0.200 g, 0.884 mmol). The reaction mixture was allowed to warm to room temperature, then heated to reflux for 3 h. Solvent was removed *in vacuo* and the resultant solid was recrystallised from MeOH to give the methyl ester **3** as yellow needles (0.180 g, 75%).

R_f (DCM:MeOH, 90:10) = 0.67; **mp** 191-193 °C, MeOH, lit.[169] 195-197 °C, MeOH; **$[\alpha]_D$ =** +10.0 (c 1.00, MeOH), lit.[169] +8.6 (c 0.80, MeOH); **IR** (neat, cm^{-1}) 3192 (OH), 1740 (C=O), 1636 (C=C), 1587 (C=C), 1541 (NO); **^{1}H NMR** δ (500 MHz, CD_3OD) 8.03 (1H, d, *J* = 2.3 Hz, Ar*H*), 7.54 (1H, dd, *J* = 8.6, 2.3 Hz, Ar*H*), 7.20 (1H, d, *J* = 8.6 Hz, Ar*H*), 4.39 (1H, dd, *J* = 7.3, 6.1 Hz, NH_2CH), 3.86 (3H, s, CO_2CH_3), 3.33-3.28 (1H, m, CH_AH_B), 3.21 (1H, dd, *J* = 14.7, 7.3 Hz, CH_AH_B); **^{13}C NMR** δ (126 MHz, CD_3OD) 168.8 (*C*O), 153.6 (Ar*C*), 137.5 (Ar*C*H), 134.5 (Ar*C*), 125.9 (Ar*C*), 125.6 (Ar*C*H), 120.4 (Ar*C*H), 55.4 (*C*H), 52.4 (CH_3), 34.6 (CH_2); ***m/z*** (ESI+, MeCN/H_2O) 241 ($[M+H]^+$, 50%), 181 ($[M-CO_2CH_3]^+$, 100).

^{1}H and ^{13}C NMR spectroscopic data are in good agreement with the literature.[169]

Methyl (2*S*)-2-acetamido-3-(4-hydroxy-3-nitrophenyl)propanoate, **4**

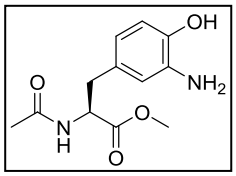


The ammonium salt **3** (0.10 g, 0.36 mmol), was dissolved in H_2O (1 mL) and cooled to 0 °C in an ice bath. NaOAc (0.72 mL, 3.6 mmol; 5 M, aq.) was added, before dropwise addition of acetic anhydride (0.05 mL, 0.53 mmol). The reaction was monitored by TLC until judged to have reached completion (~30 min). The aqueous mixture was extracted with DCM (3 × 5 mL), the combined organics were dried and solvent was removed *in vacuo*. Purification was performed by column chromatography (dry loaded, EtOAc:Hexane, 67:33) to give the protected amino acid derivative **4** as a yellow oil (69 mg, 68%).

R_f (EtOAc:Hexane, 67:33) = 0.32; **$[\alpha]_D$ =** +12.4 (c 0.10, DMSO); **IR** (neat, cm^{-1}) 3285 (OH), 1742 (C=O), 1634 (C=O), 1535 (NO); **^{1}H NMR** δ (500 MHz, $CDCl_3$) 10.50 (1H, s, O*H*), 7.87 (1H, d, *J* = 2.2 Hz, Ar*H*), 7.37 (1H, dd, *J* = 8.6, 2.2 Hz, Ar*H*), 7.12 (1H, d, *J* = 8.6 Hz, Ar*H*), 6.03 (1H, d, *J* = 7.4 Hz, CON*H*), 4.89 (1H, dt, *J* = 7.4, 5.7 Hz, C*H*), 3.79 (3H, s, CO_2CH_3), 3.22 (1H, dd, *J* = 14.1, 5.8 Hz, CH_AH_B), 3.09 (1H, dd, *J* = 14.1, 5.5 Hz, CH_AH_B), 2.04 (3H, s, CH_3CO); **^{13}C NMR** δ (126 MHz, $CDCl_3$) 171.6 (*C*O), 169.7 (*C*O), 154.2 (Ar*C*), 138.6 (Ar*C*H), 133.4 (Ar*C*), 128.5 (Ar*C*), 125.2 (Ar*C*H), 120.2 (Ar*C*H), 53.1 (*C*H), 52.7 (OCH_3), 36.8 (CH_2), 23.2 (CH_3); ***m/z*** (ESI+, $MeCN/H_2O$) 283 ($[M+H]^+$, 100%).

^{1}H and ^{13}C NMR spectroscopic data are in good agreement with the literature.[170]

Methyl (2*S*)-3-(3-amino-4-hydroxyphenyl)-2-acetamidopropanoate, **5**



Method 1: Nitrotyrosine derivative **4** (283 mg, 1.00 mmol) was dissolved in EtOH (3.8 mL) and water (0.2 mL). Iron powder (168 mg, 3.00 mmol) and anhydrous $CaCl_2$ (111 mg, 1.00 mmol) were added and the reaction mixture was heated to 60 °C for ~1 h. Iron was removed by filtration and the filtrate was concentrated *in vacuo*. The residue was purified by column chromatography (dry loaded, DCM:MeOH, 95:5) to give the aminotyrosine derivative **5** as a brown oil (156 mg, 62%).

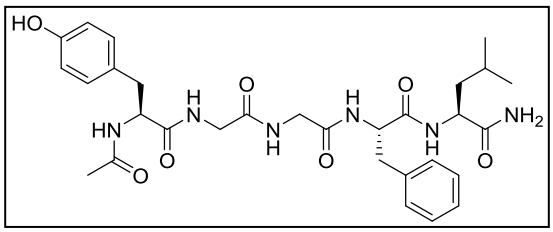
Method 2: Nitrotyrosine derivative **4** (220 mg, 0.779 mmol) was dissolved in DCM (7 mL) and 10% Pd/C (8 mg) was added. The reaction was stirred under a H_2 atmosphere for 48 h, with monitoring by TLC. The reaction mixture was filtered through celite, which was washed with EtOAc. The eluate was concentrated *in vacuo* and the product was purified by column chromatography (dry-loaded, DCM:MeOH, 95:5) to give the aminotyrosine derivative **5** as a brown oil (51 mg, 23%).

Method 3: Nitrotyrosine derivative **4** (2.136 g, 7.57 mmol) was dissolved in water (80 mL) and $Na_2S_2O_4$ (18.2 g, 88.9 mmol) was added. The reaction mixture was stirred at rt for 90 min and monitored by TLC. The aqueous mixture was lyophilised, then redissolved in MeOH and filtered to remove the salt. The eluate was concentrated *in vacuo* and purified by column chromatography (dry-loaded, DCM:MeOH, 95:5) to give the aminotyrosine derivative **5** as a brown oil (1.317 g, 69%).

R_f (DCM:MeOH, 90:10) = 0.46; **$[\alpha]_D$ =** +46.7 (c 0.30, DMSO); **IR** (neat, cm^{-1}) 3291-3177 (NH,OH), 1734 (C=O), 1649 (C=O); **^{1}H NMR** δ (500 MHz, CD_3OD) 6.63 (1H, d, *J* = 8.0 Hz,

Ar*H*), 6.61 (1H, d, *J* = 2.1 Hz, Ar*H*), 6.43 (1H, dd, *J* = 8.0, 2.1 Hz, Ar*H*), 4.59 (1H, dd, *J* = 8.4, 5.9 Hz, C*H*), 3.69 (3H, s, CO_2CH_3), 2.96 (1H, dd, *J* = 13.9, 5.9 Hz, CH_AH_B), 2.81 (1H, dd, *J* = 13.9, 8.4 Hz, CH_AH_B), 1.94 (3H, s, CH_3); **^{13}C NMR** δ (126 MHz, CD_3OD) 172.5 (*C*O), 171.8 (*C*O), 144.1 (Ar*C*), 134.9 (Ar*C*), 128.0 (Ar*C*), 119.2 (Ar*C*H), 116.7 (Ar*C*H), 114.1 (Ar*C*H), 54.3 (*C*H), 51.2 (OCH_3), 36.6 (CH_2), 20.9 (CH_3); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 253.1180, $C_{12}H_{17}N_2O_4$ requires 253.1183.

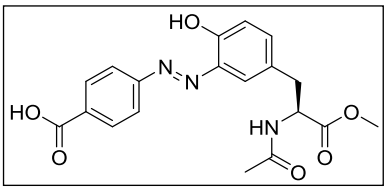
Ac-Tyr-Gly-Gly-Phe-Leu-NH_2, **6**



Ac-Leu-enkephalin-NH_2 **6** was synthesised by solid phase peptide synthesis, according to the method described in Section 5.3. The lyophilised peptide, was a colourless amorphous solid (96% purity by RP-HPLC at 210 nm, 56% isolated yield).

R_t (RP-HPLC, H_2O:MeCN, Method 1) = 10.04 min; **^{1}H NMR** δ (500 MHz, DMSO-d_6) 9.14 (1H, s, Tyr(ArOH)), 8.21 (1H, t, *J* = 5.7 Hz, Gly(N*H*)), 8.07-8.01 (2H, m, Tyr(N*H*), Phe(N*H*)), 7.98-7.91 (2H, m, Gly(N*H*), Leu(N*H*)), 7.27-7.23 (4H, m, Phe(Ar*H*)), 7.21-7.15 (1H, m, Phe(Ar*H*)), 7.07 (1H, s, $CONH_AH_B$), 7.02 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 6.95 (1H, s, $CONH_AH_B$), 6.64 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 4.51 (1H, td, *J* = 9.3, 4.6 Hz, Phe(α-C*H*)), 4.40 (1H, ddd, *J* = 9.8, 8.0, 4.5 Hz, Tyr(α-C*H*)), 4.23-4.18 (1H, m, Leu(α-C*H*)), 3.74-3.70 (2H, m, Gly(α-CH_AH_B), Gly(α-CH_AH_B)), 3.68-3.60 (2H, m, Gly(α-CH_AH_B), Gly(α-CH_AH_B)), 3.04 (1H, dd, *J* = 13.8, 4.6 Hz, Phe(CH_AH_B)), 2.91 (1H, dd, *J* = 13.9, 4.5 Hz, Tyr(CH_AH_B)), 2.80 (1H, dd, *J* = 13.8, 9.3 Hz, Phe(CH_AH_B)), 2.64 (1H, dd, *J* = 13.9, 9.8 Hz, Tyr(CH_AH_B)), 1.77 (3H, s, CH_3CONH), 1.63-1.52 (1H, m, Leu(γ-C*H*)), 1.49-1.46 (2H, m, Leu(β-CH_2)), 0.89 (3H, d, *J* = 6.6 Hz, Leu(CH_3)), 0.84 (3H, d, *J* = 6.5 Hz, Leu(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 174.4 (*C*O), 172.5 (*C*O), 171.1 (*C*O), 169.8 (*C*O), 169.6 (*C*O), 169.1 (*C*O), 156.2 (Ar*C*), 138.2 (Ar*C*), 130.5 (2Ar*C*H), 129.7 (2Ar*C*H), 128.6 (Ar*C*), 128.5 (2Ar*C*H), 126.7 (Ar*C*H), 115.3 (2Ar*C*H), 55.0 (*C*H), 54.5 (*C*H), 51.5 (*C*H), 42.6 (CH_2), 42.4 (CH_2), 41.3 (CH_2), 37.8 (CH_2), 37.1 (CH_2), 24.7 (*C*H), 23.5 (CH_3), 23.0 (CH_3), 22.1 (CH_3); **HRMS** (ESI+, MeCN/H_20) $[M+H]^+$ found 597.3034 $C_{30}H_{41}N_6O_7$ requires 597.3031.

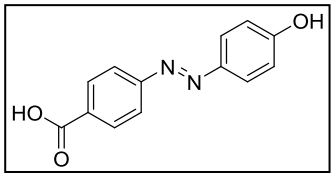
4-[(*E*)-2-{5-[(2*S*)-2-Acetamido-3-methoxy-3-oxopropyl]-2-hydroxyphenyl}diazen-1-yl]benzoic acid, **7**



4-Aminobenzoic acid (74.0 mg, 0.540 mmol) was dissolved in HCl (0.07 mL; conc.) and H_2O (1.5 mL) and cooled to 0 °C in an ice bath. $NaNO_2$ (61.0 mg, 0.884 mmol) in H_2O (0.5 mL) was added to the reaction mixture and stirred for 20 min whereupon a pale yellow colour developed. Tyrosine derivative **2** (100 mg, 0.421 mmol), NaOH (15.0 mg, 0.378 mmol) and K_2CO_3 (76.0 mg, 0.547 mmol) were added and the reaction mixture was allowed to warm to room temperature. The reaction was monitored by TLC until judged to have reached completion (~1 h). HCl (1 mL; 1 M, aq.) was added and the product was removed by filtration. The resultant solids were purified by column chromatography (dry loaded, DCM:MeOH:acetic acid, 90:5:5) to give the azobenzene derivative **7** as an orange solid (88.0 mg, 42%).

R_f (DCM:MeOH, 90:10) = 0.28; **mp** 239–241 °C; **IR** (neat, cm^{-1}) 3254 (OH), 1734 (CO), 1684 (CO), 1653 (CO), 1593 (C=C), 1541 (C=C); **^{1}H NMR** δ (500 MHz, CD_3OD) 8.23 (2H, d, *J* = 8.1 Hz, Ar*H*), 8.03 (2H, d, *J* = 8.1 Hz, Ar*H*), 7.82 (1H, d, *J* = 2.3 Hz, Ar*H*), 7.32 (1H, dd, *J* = 8.5, 2.3 Hz, Ar*H*), 7.01 (1H, d, *J* = 8.5 Hz, Ar*H*), 4.77-4.70 (1H, m, C*H*), 3.74 (3H, s, OCH_3), 3.22 (1H, dd, *J* = 14.0, 5.7 Hz, CH_AH$_B$), 3.03 (1H, dd, *J* = 14.0, 8.8 Hz, CH$_A$$H_B$), 1.95 (3H, s, C$H_3$CONH); **^{13}C NMR** δ (126 MHz, CD_3OD) 172.1 (*C*O), 171.8 (*C*O), 153.8 (*C*O), 153.8 (Ar*C*), 153.4 (Ar*C*), 152.3 (Ar*C*), 137.7 (Ar*C*), 134.8 (Ar*C*H), 130.7 (2Ar*C*H), 130.1 (Ar*C*H), 128.8 (Ar*C*), 121.9 (2Ar*C*H), 117.9 (Ar*C*H), 53.9 (*C*H), 51.4 (O*C*H$_3$), 36.0 (*C*H$_2$), 20.9 (*C*H$_3$); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 386.1353, $C_{19}H_{20}N_3O_6$ requires 386.1347.

4-[(*E*)-2-(4-Hydroxyphenyl)diazen-1-yl]benzoic acid, **8**

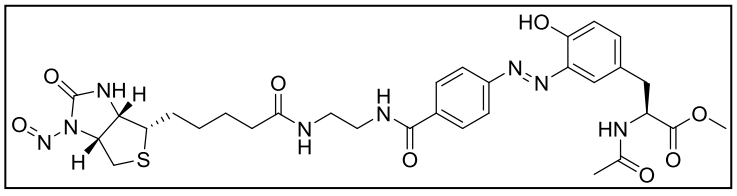


4-Aminobenzoic acid (123 mg, 0.900 mmol) was dissolved in HCl (0.12 mL; conc.) and water (2.4 mL) and cooled to 0 °C. Sodium nitrite (104 mg, 1.50 mmol) was dissolved in water (0.5 mL) and added to the reaction mixture. The mixture was stirred for 20 min whereupon a pale yellow colour developed. A mixture of phenol (67 mg, 0.720 mmol), NaOH (26 mg, 0.642 mmol) and K_2CO_3 (133 mg, 0.960 mmol) in water (2 mL) was added dropwise. The reaction mixture was allowed to warm to room temperature and stirred for 4 h. The reaction mixture was next acidified with HCl (~2 mL; 1 M, aq.) and the product was removed by filtration. The resultant solids were purified by column chromatography (dry loaded, DCM:MeOH, 90:10) to give the azobenzene derivative **8** as an orange solid (144 mg, 83%).

R_f (DCM:MeOH, 95:5) = 0.25; **mp** 276-279 °C, lit.[66] 275-277 °C; **IR** (neat, cm^{-1}) 3231 (OH), 1678 (C=O); **^{1}H NMR** δ (500 MHz, CD_3OD) 8.18 (2H, d, *J* = 8.6 Hz, Ar*H*), 7.91 (2H, d, *J* = 8.6 Hz, Ar*H*), 7.89 (2H, d, *J* = 8.9 Hz, Ar*H*) 6.96 (2H, d, *J* = 8.9 Hz, Ar*H*); **^{13}C NMR** (126 MHz, CD_3OD) 161.5 (*C*O), 155.4 (Ar*C*), 146.2 (Ar*C*), 132.9 (Ar*C*), 131.9 (Ar*C*), 130.4 (2Ar*C*H), 125.0 (2Ar*C*H), 121.8 (2Ar*C*H), 115.5 (2Ar*C*H); ***m/z*** (ESI+, MeCN/H_2O) 243 ($[M+H]^+$, 100%).

^{1}H and ^{13}C NMR spectroscopic data are in good agreement with the literature.[66]

Methyl (2*S*)-3-{3-[(*E*)-2-{4-[(2-{5-[(3a*S*,4*S*,6a*R*)-1-nitroso-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]pentanamido}ethyl)carbamoyl]phenyl}diazen-1-yl]-4-hydroxyphenyl}-2-acetamidopropanoate, **13**

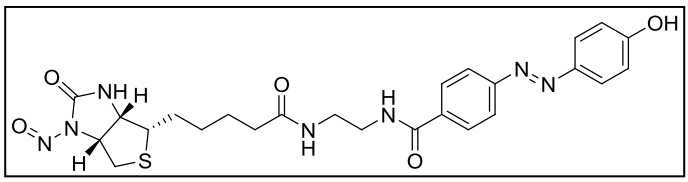


Boc-protected aniline **32** (70.0 mg, 0.138 mmol) was dissolved in DCM (2 mL) and TFA (0.150 mL, 1.96 mmol) was added. The reaction was stirred at room temperature and monitored by TLC until judged to have reached completion (~18 h). Solvent was removed *in vacuo*, and the residue was redissolved in HCl (0.3 mL; 6 N, aq.) and cooled to 0 °C in an ice bath. $NaNO_2$ (24.0 mg, 0.345 mmol) was added and the reaction mixture was stirred for 20 min as a pale yellow colour developed. Tyrosine derivative **2** (16 mg, 0.067 mmol) was dissolved in $NaHCO_3$ (2 mL; sat. aq.) and cooled to 0

°C. The diazonium salt mixture was added dropwise to the tyrosine-containing mixture and the reaction mixture was maintained at >pH 7 by addition of solid $NaHCO_3$, with monitoring by pH paper. The reaction mixture was allowed to reach room temperature and monitored by TLC until judged to have reached completion (~15 min). The reaction mixture was acidified with HCl (4 mL; 1 M, aq.) and extracted with EtOAc (3 × 5 mL). The combined organics were dried and concentrated *in vacuo*. Purification was performed by column chromatography (dry-loaded, DCM:MeOH, 90:10) to give the azobenzene derivative **13** as an orange solid (28 mg, 64%).

R_f (DCM:MeOH, 90:10) = 0.34; **mp** 178-181 °C; **$[\alpha]_D$ =** +115.0 (c 0.20, DMSO); **IR** (neat, cm^{-1}) 3277 (OH), 1744 (CO), 1643 (CO), 1535 (C=C); **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 10.92 (1H, s, OH), 8.80 (1H, s, Biotin-NH), 8.66 (1H, t, *J* = 5.5 Hz, N*H*CO), 8.36 (1H, d, *J* = 7.8 Hz, CH_3CON*H*), 8.09-8.01 (4H, m, Ar*H*), 7.96 (1H, t, *J* = 5.7 Hz, CON*H*), 7.61 (1H, d, *J* = 2.1 Hz, Ar*H*), 7.32 (1H, dd, *J* = 8.4, 2.1 Hz, Ar*H*), 7.02 (1H, d, *J* = 8.4 Hz, Ar*H*), 4.82 (1H, t, *J* = 6.3 Hz, C*H*CH_2S), 4.47 (1H, td, *J* = 8.0, 5.7 Hz, C*H*CO_2CH_3), 4.25 (1H, dd, *J* = 6.8, 4.5 Hz, C*H*CHS), 3.62 (3H, s, OC*H*$_3$), 3.37-3.33 (2H, m, C*H*$_2$), 3.26 (2H, dd, *J* = 12.4, 6.2 Hz, CONHC*H*$_2$), 3.22-3.15 (1H, m, C*H*S), 3.02 (1H, dd, *J* = 13.8, 5.7 Hz, ArC*H*$_A$H$_B$), 2.95 (1H, dd, *J* = 13.3, 5.8 Hz, C*H*$_X$H$_Y$S), 2.89 (1H, dd, *J* = 13.8, 9.2, ArCH$_A$*H*$_B$), 2.67 (1H, d, *J* = 13.3 Hz, CH$_X$*H*$_Y$S), 2.10 (2H, t, *J* = 7.0 Hz, C*H*$_2$CONH), 1.81 (3H, s, C*H*$_3$CONH), 1.75-1.64 (1H, m, SCHC*H*$_M$H$_N$), 1.62-1.44 (3H, m, SCHCH$_M$*H*$_N$, C*H*$_2$), 1.42-1.21 (2H, m, C*H*$_2$); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 172.8 (*C*O), 172.6 (*C*O), 169.8 (*C*O), 166.1 (*C*O), 154.4 (*C*O), 154.2 (Ar*C*), 153.5 (Ar*C*), 138.7 (Ar*C*), 136.8 (Ar*C*), 135.5 (Ar*C*H), 129.2 (Ar*C*), 128.9 (2Ar*C*H), 122.9 (2Ar*C*H), 122.6 (Ar*C*H), 118.7 (Ar*C*H), 60.0 (*C*H), 59.0 (*C*H), 54.4 (*C*H), 54.1 (*C*H), 52.3 (O*C*H$_3$), 39.9 (*C*H$_2$), 38.7 (*C*H$_2$), 36.3 (*C*H$_2$), 35.7 (*C*H$_2$), 35.0 (*C*H$_2$), 28.7 (*C*H$_2$), 28.1 (*C*H$_2$), 25.6 (*C*H$_2$), 22.7 (*C*H$_3$); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 683.2605, $C_{31}H_{39}N_8O_8S$ requires 683.2606.

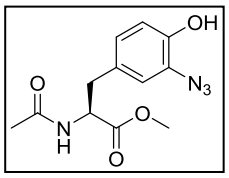
5-[(3a*S*,4*S*,6a*R*)-1-Nitroso-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]-*N*-[2-({4-[(*E*)-2-(4-hydroxyphenyl)diazen-1-yl]phenyl}formamido)ethyl]pentanamide, **14**



Boc-protected aniline **32** (70.0 mg, 0.138 mmol) was dissolved in DCM (2 mL) at room temperature. TFA (0.15 mL, 1.96 mmol) was added and the reaction was monitored by TLC until judged to have reached completion (~18 h). Solvent was removed *in vacuo* before addition of HCl (0.3 mL; 6 N, aq.) and the reaction mixture was cooled to 0 °C in an ice bath. $NaNO_2$ (24.0 mg, 0.345 mmol) was added and the reaction mixture was stirred for 20 min as a pale yellow colour developed. Phenol (6.0 mg, 0.064 mmol) was dissolved in $NaHCO_3$ (2 mL; sat. aq.) and cooled to 0 °C. The diazonium salt mixture was added dropwise to the phenol-containing solution and the reaction mixture was maintained at >pH 7 by addition of solid $NaHCO_3$, with monitoring by pH paper. Once addition was complete, the reaction mixture was allowed to reach room temperature and was monitored by TLC until completion (~15 min). The reaction mixture was acidified by addition of HCl (4 mL; 1 M, aq.) and extracted with EtOAc (3 × 5 mL). The combined organics were dried and concentrated *in vacuo*. Purification was performed by column chromatography (dry-loaded, DCM:MeOH, 95:5) to give the azobenzene derivative **14** as an orange solid (10.0 mg, 31%).

R_f (DCM:MeOH, 95:5) = 0.26; **mp** 174-176 °C; **IR** (neat, cm^{-1}) 3294 (OH), 1749 (C=O), 1634 (C=O), 1589 (C=C), 1539 (C=C); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 10.40 (1H, br s, O*H*), 8.80 (1H, s, Biotin-N*H*), 8.63 (1H, t, *J* = 5.6 Hz, N*H*CO), 8.02 (2H, d, *J* = 8.6 Hz, 2Ar*H*), 7.96 (1H, t, *J* = 5.7 Hz, CON*H*), 7.87 (2H, d, *J* = 8.6 Hz, 2Ar*H*), 7.84 (2H, d, *J* = 8.8 Hz, 2Ar*H*), 6.97 (2H, d, *J* = 8.8 Hz, 2Ar*H*), 4.82 (1H, ddd, *J* = 7.1, 5.9, 0.9 Hz, C*H*CH_2S), 4.24 (1H, dd, *J* = 7.1, 4.3 Hz, C*H*CHS), 3.37-3.31 (2H, m, CH_2), 3.26 (2H, dt, *J* = 6.9, 6.4 Hz, CONHCH_2), 3.20-3.15 (1H, m, C*H*S), 2.94 (1H, dd, *J* = 13.3, 5.9 Hz, C$H_A$$H_B$S), 2.67 (1H, d, *J* = 13.3 Hz, C$H_A$$H_B$S), 2.11 (2H, t, *J* = 7.4 Hz, CH_2CO), 1.75-1.64 (1H, m, SCHC$H_X$$H_Y$), 1.62-1.45 (3H, m, SCHC$H_X$$H_Y$, C$H_2$), 1.42-1.12 (2H, m, C$H_2$); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 172.9 (*C*O), 166.2 (*C*O), 161.9 (*C*O), 154.2 (Ar*C*), 154.0 (Ar*C*), 145.8 (Ar*C*), 136.2 (Ar*C*), 128.9 (2Ar*C*H), 125.6 (2Ar*C*H), 122.3 (2Ar*C*H), 116.5 (2Ar*C*H), 60.0 (*C*H), 59.0 (*C*H), 54.4 (*C*H), 39.9 (*C*H_2), 38.7 (*C*H_2), 35.7 (*C*H_2), 35.0 (*C*H_2), 28.6 (*C*H_2), 28.1 (*C*H_2), 25.6 (*C*H_2); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 540.2043, $C_{25}H_{30}N_7O_5S$ requires 540.2029.

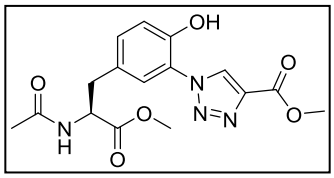
Methyl (2*S*)-3-(3-azido-4-hydroxyphenyl)-2-acetamidopropanoate, **15**



3-Aminotyrosine derivative **5** (88 mg, 0.35 mmol) was dissolved in HCl (1.2 mL; 0.5 N, aq.) and cooled to 0 °C in an ice bath in the dark. A solution of $NaNO_2$ (36 mg, 0.52 mmol) in water (0.4 mL) was added dropwise and the mixture was stirred for 10 min. A solution of NaN_3 (57 mg, 0.88 mmol) in water (0.4 mL) was added and the reaction mixture was stirred at 0 °C and monitored by TLC until judged to have reached completion (~6 h). The desired product was extracted in $CHCl_3$ (3 × 5 mL) and solvent was removed *in vacuo*. Purification was performed by column chromatography (dry-loaded, EtOAc:hexane, 3:1) to give the 3-azidotyrosine derivative **15** as a brown oil (56 mg, 67%).

R_f (DCM:MeOH, 95:5) = 0.26; **IR** (neat, cm^{-1}) 3265 (OH), 2112 (N_3), 1732 (C=O), 1651 (C=O), 1512 (C=C); **^{1}H NMR** δ (500 MHz, CD_3OD) 6.84 (1H, dd, *J* = 8.2, 2.0 Hz, Ar*H*), 6.82 (1H, d, *J* = 2.0 Hz, Ar*H*), 6.77 (1H, d, *J* = 8.2 Hz, Ar*H*), 4.62 (1H, dd, *J* = 8.7, 5.7 Hz, C*H*), 3.71 (3H, s, OCH_3), 3.05 (1H, dd, *J* = 14.0, 5.7 Hz, CH_AH$_B$), 2.86 (1H, dd, *J* = 14.0, 8.7 Hz, CH$_A$$H_B$), 1.94 (3H, s, C$H_3$CONH); **^{13}C NMR** δ (126 MHz, CD_3OD) 172.1 (*C*O), 171.7 (*C*O), 148.7 (Ar*C*), 128.7 (Ar*C*), 126.3 (Ar*C*), 126.2 (Ar*C*H), 120.6 (Ar*C*H), 116.0 (Ar*C*H), 54.0 (*C*H), 51.3 (OCH_3), 36.1 (CH_2), 20.9 (CH_3); ***m/z*** (ESI+, MeCN/H_2O) 301 ($[M+Na]^+$, 3%), 282 ($[M-N_3+CO_2H+H]^+$, 100), 279 ($[M+H]^+$, 1).

Methyl 1-{5-[(2*S*)-2-acetamido-3-methoxy-3-oxopropyl]-2-hydroxyphenyl}-1*H*-1,2,3-triazole-4-carboxylate, **16**

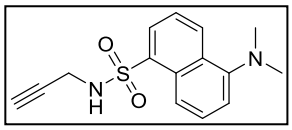


Azide **15** (66 mg, 24 μmol) was dissolved in *t*-amyl alcohol (1.4 mL) and water (0.7 mL). $CuSO_4·5H_2O$ (6.0 mg, 24 μmol) and sodium ascorbate (9.5 mg, 48 μmol) were added and the reaction mixture was stirred at room temperature for 30 min. Methyl propiolate (24 μl, 270 μmol) was added and the reaction was stirred for 16 h. Solvent was removed *in vacuo* and purification was performed by column chromatography (DCM:MeOH, 95:5) to give the triazole **16** (76 mg, 89%).

R_f (EtOAc:Hexane, 4:1) = 0.22; **IR** (neat, cm^{-1}) 3543 (OH), 1734 (C=O), 1719 (C=O), 1638 (C=O), 1533 (C=C), 1518 (C=C); **^{1}H NMR** δ (500 MHz, $CDCl_3$) 9.60 (1H, s, O*H*), 8.81 (1H, s, triazole-C*H*), 7.57 (1H, d, *J* = 2.1 Hz, Ar*H*), 7.13 (1H, dd, *J* = 8.4, 2.1 Hz, Ar*H*), 7.04 (1H, d, *J* = 8.4 Hz, Ar*H*), 6.53 (1H, d, *J* = 8.3 Hz, CON*H*), 4.97-4.93 (1H, m, CONHC*H*), 4.00 (3H, s, triazole-CO_2CH_3), 3.80 (3H, s, CO_2CH_3), 3.23 (1H, dd, *J* = 14.1, 5.2 Hz, CH_AH$_B$), 3.00 (1H,

dd, *J* = 14.1, 7.7 Hz, CH_AH_B), 2.03 (3H, s, CH_3CONH); **^{13}C NMR** δ (500 MHz, $CDCl_3$) 171.9 (*C*O), 170.8 (*C*O), 161.2 (*C*O), 147.7 (Ar*C*), 139.5 (Ar*C*), 130.9 (Ar*C*H), 128.4 (Ar*C*), 127.8 (triazole-*C*H), 123.2 (Ar*C*H), 123.1 (Ar*C*), 118.6 (Ar*C*H), 53.3 (*C*H), 52.7 (OCH_3), 52.4 (OCH_3), 37.4 (CH_2), 23.0 (CH_3CONH); **HRMS** (ESI+, $MeCN/H_2O$) $[M+H]^+$ found 363.1305, $C_{16}H_{19}N_4O_6$ requires 363.1299.

5-(Dimethylamino)-*N*-(prop-2-yn-1-yl)naphthalene-1-sulphonamide, **17**

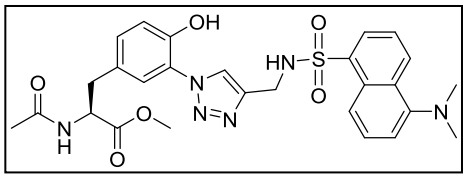


Dansyl chloride (0.10 g, 0.37 mmol) was dissolved in DCM (1 mL) and stirred at 0 °C in the dark. NEt_3 (52 µl, 1.1 mmol) and propargylamine (72 µl, 1.1 mmol) were added and the reaction mixture was stirred at 0 °C for 1 h. Solvent was removed *in vacuo* and purification was performed by column chromatography (diethyl ether:hexane, 1:1) to give the alkyne **17** as a yellow solid (83 mg, 77%).

R_f (Et_2O:Hexane, 1:1) = 0.39; **mp** 89-91 °C, lit.[171] 93-95 °C; **IR** (neat, cm^{-1}) 3285 (NH), 1670 (C=C), 1574 (C=C); **^{1}H NMR** δ (500 MHz, $CDCl_3$) 8.59 (1H, ddd, *J* = 8.5, 1.2, 1.0 Hz, Ar*H*), 8.33-8.25 (2H, m, Ar*H*), 7.61 (1H, dd, *J* = 8.6, 7.6 Hz, Ar*H*), 7.55 (1H, dd, *J* = 8.5, 7.3 Hz, Ar*H*), 7.22 (1H, dd, *J* = 7.6, 0.9 Hz, Ar*H*), 4.79 (1H, t, *J* = 6.0 Hz, SO_2NH), 3.80 (2H, dd, *J* = 6.1, 2.6 Hz, $CH_2C{\equiv}CH$), 2.92 (6H, s, $N(CH_3)_2$), 1.94 (1H, t, *J* = 2.5 Hz, C≡*H*); **^{13}C NMR** δ (126 MHz, $CDCl_3$) 152.1 (Ar*C*), 134.2 (Ar*C*), 130.9 (Ar*C*H), 130.0 (Ar*C*H), 129.9 (Ar*C*), 129.8 (Ar*C*), 128.6 (Ar*C*H), 123.2 (Ar*C*H), 118.5 (Ar*C*H), 115.3 (Ar*C*H), 77.8 (*C*), 72.7 (*C*H), 45.4 ($2CH_3$), 33.1 (CH_2); ***m/z*** (ESI+, $MeCN/H_2O$) 289 ($[M+H]^+$, 100%).

^{1}H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[171]

Methyl (2*S*)-3-(3-{4-[5-(dimethylamino)naphthalene-1-sulfonamidomethyl]-1*H*-1,2,3-triazol-1-yl}-4-hydroxyphenyl)-2-acetamidopropanoate, **18**

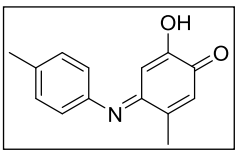


3-Azidotyrosine derivative **15** (50 mg, 0.18 mmol) was dissolved in *t*-amyl alcohol (1.4 mL) and water (0.7 mL) and stirred at room temperature in the dark. $CuSO_4{\cdot}5H_2O$ (4.5 mg, 0.018 mmol) and sodium ascorbate (7.5 mg, 0.038 mmol) were added and the reaction mixture was stirred for 20 min. Dansyl alkyne **17** (62 mg, 0.22 mmol) was added and the reaction mixture was stirred at room temperature for 18 h. The desired product was extracted in DCM (3 × 5 mL), dried and purified

by column chromatography (dry-loaded, DCM:MeOH, 95:5) to give the triazole **18** as a yellow oil (85 mg, 83%).

R_f (DCM:MeOH, 95:5) = 0.38; **IR** (neat, cm^{-1}) 3071 (OH), 1740 (C=O), 1657 (C=O), 1517 (C=C); **^{1}H NMR** δ (500 MHz, DMSO-*d*$_6$) 10.36 (1H, s, O*H*), 8.50 (1H, t, *J* = 6.0 Hz, N*H*SO_2), 8.42 (1H, d, *J* = 8.5 Hz, Ar*H*), 8.35-8.28 (2H, m, CON*H*, Ar*H*), 8.13 (1H, dd, *J* = 7.3, 1.2 Hz, Ar*H*), 7.97 (1H, s, triazole-C*H*), 7.60-7.56 (2H, m, Ar*H*), 7.35 (1H, d, *J* = 2.2 Hz, Tyr(Ar*H*)), 7.22 (1H, dd, *J* = 7.6, 0.9 Hz, Ar*H*), 7.16 (1H, dd, *J* = 8.4, 2.2 Hz, Tyr(Ar*H*)), 6.99 (1H, d, *J* = 8.4 Hz, Tyr(Ar*H*)), 4.44 (1H, ddd, *J* = 9.1, 7.7, 5.6 Hz, Tyr(α-C*H*)), 4.21 (2H, d, *J* = 5.8 Hz, C*H*$_2$NHSO_2), 3.62 (3H, s, OC*H*$_3$), 2.99 (1H, dd, *J* = 13.8, 5.6 Hz, C*H*$_A$H$_B$), 2.86 (1H, dd, *J* = 13.8, 9.1 Hz, CH$_A$*H*$_B$), 2.77 (6H, s, N(C*H*$_3$)$_2$), 1.82 (3H, s, C*H*$_3$CONH); **^{13}C NMR** δ (500 MHz, $CDCl_3$) 172.6 (*C*O), 169.8 (*C*O), 151.8 (Ar*C*), 148.2 (Ar*C*), 143.8 (Ar*C*), 136.5 (Ar*C*), 130.8 (Ar*C*H), 129.9 (Ar*C*H), 129.5 (Ar*C*), 129.5 (Ar*C*), 128.9 (Ar*C*), 128.8 (Ar*C*H), 128.3 (Ar*C*H), 125.5 (Ar*C*H), 124.8 (triazole-*C*H), 124.4 (Ar*C*), 124.0 (Ar*C*H), 119.6 (Ar*C*H), 117.4 (Ar*C*H), 115.5 (Ar*C*H), 54.2 (*C*H), 52.3 (*C*H$_3$), 45.5 (2*C*H$_3$), 38.3 (*C*H$_2$), 36.2 (*C*H$_2$), 22.7 (*C*H$_3$); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 567.2025, $C_{27}H_{31}N_6O_6S$ requires 567.2020.

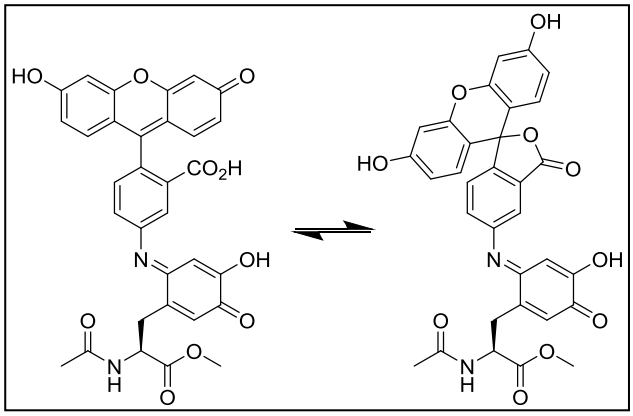
2-hydroxy-5-methyl-4-[(4-methylphenyl)imino]cyclohexa-2,5-dien-1-one, **24**



2-Amino-4-methylphenol (12.4 mg, 0.1 mmol) in MeCN (200 µL) and 4-methylaniline (10.7 mg, 0.1 mmol) in MeCN (200 µL) were added to phosphate buffer (90 mL, 10 mM, pH 6.5). K_3FeCN_6 (330 mg, 1 mmol) was added in H_2O (10 mL) and the reaction mixture was stirred at rt for 30 min, with analysis by HPLC. The reaction mixture was lyophilised and then redissolved in $CHCl_3$ and filtered to remove the Fe species. The product was purified by preparative RP-HPLC (System 4) using an isocratic solvent system (H_2O + 0.1% TFA: MeCN + 0.1% TFA, 56:44). The collected fractions were lyophilised to give the oxidative coupling product **24** as a red/brown amorphous solid.

R_t (HPLC, Method 1): 10.29 min (dimer by-product at 10.02 min); **^{1}H NMR** δ (500 MHz, DMSO-*d*$_6$) δ 8.57 (1H, br s, O*H*), 6.93 – 6.87 (2H, m, 2Ar*H*), 6.57 – 6.56 (2H, m, 2Ar*H*), 6.56 – 6.52 (2H, m, 2Ar*H*), 2.16 (3H, s, C*H*$_3$), 1.97 (3H, s, C*H*$_3$); **^{13}C NMR** δ (126 MHz, DMSO-*d*$_6$) 145.2 (Ar*C*), 143.7 (Ar*C*), 141.7 (Ar*C*), 132.9 (Ar*C*), 129.7 (2Ar*C*H), 125.9 (Ar*C*), 122.8 (Ar*C*), 118.1 (Ar*C*H), 114.7 (2Ar*C*H), 111.9 (Ar*C*H), 20.6 (*C*H$_3$), 17.5 (C*H*$_3$); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 228.1017, $C_{14}H_{14}NO_2$ requires 228.1019.

5-({2-[(2*S*)-2-acetamido-3-methoxy-3-oxopropyl]-5-hydroxy-4-oxocyclohexa-2,5-dien-1-ylidene}amino)-2-(6-hydroxy-3-oxo-3*H*-xanthen-9-yl)benzoic acid, **26**

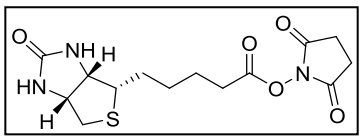


Ac-Tyr(3-NH_2)-OMe **5** (100 µL, 100 mM in DMSO:buffer, 1:1) and Fluoresceinamine (100 µL, 1 M in DMSO) were added to phosphate buffer (10 mL, 10 mM, pH 6.5). K_3FeCN_6 (200 µL, 500 mM in buffer) was added and the reaction mixture was stirred at rt for 2 h, and monitored by HPLC. The reaction mixture was concentrated *in vacuo* and desalted using a prepacked C_{18} silica column (Supelco DSC-18), with elution in MeOH. The product was concentrated *in vacuo*, then suspended in H_2O/EtOAc. The aqueous layer was acidified by addition of HCl (1 M, aq.) and the product was extracted in EtOAc. Finally, the desired product was purified by preparative RP-HPLC (System 4) using an isocratic solvent system (H_2O + 0.1%: MeCN + 0.1% TFA, 65:35). The collected fractions were lyophilised to give the oxidative coupling product **26** as a red amorphous solid.

R_t (HPLC, Method 1): 9.80 min (dimer by-product at 8.61 min); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 10.78 (1H, s, ArO*H*), 10.13 (2H, s, 2ArO*H*), 8.35 (1H, d, *J* = 8.4 Hz, N*H*), 7.39 – 7.30 (2H, m, 2Ar*H*), 7.25 (1H, dd, *J* = 8.1, 1.9 Hz, Ar*H*), 6.76 – 6.68 (4H, m, 4Ar*H*), 6.62-6.56 (3H, m, 3Ar*H*), 6.16 (1H, s, Ar*H*), 4.85 – 4.74 (1H, m, Tyr(α-C*H*)), 3.64 (3H, s, OCH_3), (CH_AH$_B$ obscured by water peak, expected ~3.25 ppm), 2.86 – 2.74 (1H, m, CH$_A$$H_B$), 1.82 (3H, s, C$H_3$CONH); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 597.1499, $C_{32}H_{25}N_2O_{10}$ requires 597.1504.

5.7 Experimental procedures for Chapter 3

2,5-Dioxopyrrolidin-1-yl 5-[(3a*S*,4*S*,6a*R*)-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]pentanoate, **27**

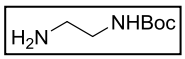


D-Biotin (1.50 g, 6.14 mmol) was stirred in hot DMF (45 mL) until dissolved. *N*-hydroxysuccinimide (0.707 g, 6.14 mmol) was added to the reaction mixture, followed by DCC (1.90 g, 9.21 mmol) and DMAP (0.075 g, 0.614 mmol) and the reaction was stirred at room temperature for 18 h. The urea by-product was removed by filtration and the filtrate was concentrated to dryness *in vacuo*. The resulting solid was triturated with diethyl ether (4 × 40 mL), then filtered, washed with propan-2-ol (40 mL) and dried *in vacuo* to give the NHS-ester, **27**, as a colourless solid (1.92 g, 92%).

R_f (DCM:MeOH, 80:20) = 0.67; **mp** 201-203 °C, lit.[172] 203-204 °C; **$[\alpha]_D$** = +50.0 (c 1.00, DMSO); **IR** (neat, cm^{-1}) 3223 (NH), 1817 (C=O), 1788 (C=O), 1746 (C=O), 1728 (C=O), 1697 (C=O); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 6.40 (1H, s, N*H*), 6.34 (1H, s, N*H*), 4.37-4.27 (1H, m, C*H*CH_2S), 4.16 (1H, ddd, *J* = 7.7, 4.5, 1.8 Hz, C*H*CHS), 3.12 (1H, ddd, *J* = 8.4, 6.4, 4.5 Hz, C*H*S), 2.89-2.76 (5H, m, 2COCH_2, C$H_A$$H_B$S), 2.68 (2H, t, *J* = 7.4 Hz, COCH_2), 2.59 (1H, d, *J* = 12.4 Hz, C$H_A$$H_B$S), 1.72-1.58 (2H, m, C$H_2$), 1.58 – 1.34 (4H, m, C$H_2$); **^{13}C NMR** (126 MHz, DMSO-d_6) δ 170.7 (2CO), 169.4 (CO), 163.1 (CO), 61.5 (CH), 59.6 (CH), 55.7 (CH), 40.4 (CH_2), 30.5 (CH_2), 28.3 (CH_2), 28.1 (CH_2), 25.9 (2CH_2), 24.8 (CH_2); ***m/z*** (ESI+, MeCN/H_2O) 364 ($[M+Na]^+$, 100%), 342 ($[M+H]^+$, 100), 227 ($[M-C_4H_4NO_3]^+$, 100).

^{1}H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[172]

tert-Butyl *N*-(2-aminoethyl)carbamate, **28**

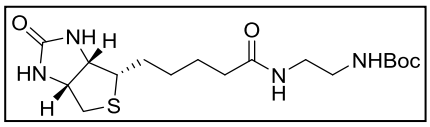


A solution of di-*tert*-butyl dicarbonate (1.91 g, 8.75 mmol) in DCM (120 mL) was added dropwise over 4-5 h to a stirred solution of ethane-1,2-diamine (3.5 mL, 52.4 mmol) in DCM (30 mL). The reaction mixture was stirred at room temperature for 18 h. Solvent was removed *in vacuo* and the remaining oil was partitioned between $NaHCO_3$ (50 mL; sat. aq.) and DCM (3 × 50 mL). The combined organics were dried and concentrated *in vacuo* to give the Boc-protected diamine **28** as a colourless oil (1.14 g, 82%). The crude product was used in a following reaction with no further purification.

R_f (DCM:MeOH, 80:20) = 0.24; **IR** (neat, cm^{-1}) 3298 (NH), 1701 (C=O); **^{1}H NMR** δ (500 MHz, $CDCl_3$) 4.94 (1H, br s, N*H*Boc) 3.22-3.19 (2H, m, CH_2NHBoc), 2.83 (2H, t, *J* = 6.3 Hz, NH_2CH_2), 1.47 (9H, s C(CH_3)$_3$); **^{13}C NMR** δ (126 MHz, $CDCl_3$) 156.3 (CO), 79.3 (C), 43.2 (CH_2), 41.8 (CH_2), 28.4 ($3CH_3$); ***m/z*** (ESI+, MeCN/H_2O) 183 ($[M+Na]^+$, 100%), 161 ($[M+H]^+$, 60).

^{1}H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[173]

tert-Butyl *N*-(2-{5-[(3a*S*,4*S*,6a*R*)-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]pentanamido}ethyl) carbamate, **29**



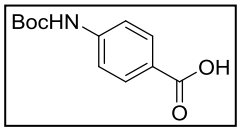
Amine **28** (0.96 g, 5.99 mmol) was dissolved in dry DMF (70 mL), before addition of biotin-NHS **27** (2.10 g, 6.15 mmol) and NEt_3 (1.56 mL, 11.2 mmol). The reaction mixture was stirred at room temperature for 18 h. Solvent was removed *in vacuo* and the resulting solid was purified by column chromatography (dry-loaded, DCM:MeOH, 90:10) to give the Boc-protected amine **29** as a colourless solid (1.87 g, 81%).

R_f (DCM:MeOH, 90:10) = 0.28; **mp** 174-176 °C; **IR** (neat, cm^{-1}) 3325 (NH), 1699 (C=O), 1684 (C=O), 1626 (C=O); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 7.76 (1H, t, *J* = 5.5 Hz, CON*H*), 6.75 (1H, t, *J* = 5.2 Hz, CON*H*), 6.39 (1H, s, Biotin-N*H*), 6.34 (1H, s, Biotin-N*H*), 4.34-4.28 (1H, m, C*H*CH_2S), 4.16-4.11 (1H, m, C*H*CHS), 3.14-3.02 (3H, m, NHCH_2, C*H*S), 3.00-2.92 (2H, m, NHCH_2), 2.83 (1H, dd, *J* = 12.4, 5.1 Hz, C$H_A$$H_B$S), 2.60 (1H, d, *J* = 12.5 Hz, CH_AH_BS), 2.05 (2H, t, *J* = 7.5 Hz, COCH_2), 1.67-1.56 (1H, m, C$H_X$$H_Y$), 1.56-1.43 (3H, m, C$H_2$, CH_XH_Y), 1.38 (9H, s, C(CH_3)$_3$), 1.35-1.25 (2H, m, CH_2); **^{13}C NMR** δ (126MHz, DMSO-d_6) 172.6 (CO), 163.2 (CO), 156.1 (CO), 78.1 (C), 61.5 (CH), 59.7 (CH), 55.9 (CH), 40.3 (CH_2), 40.2 (CH_2),

39.1 (CH_2), 35.7 (CH_2), 28.7 ($3CH_3$), 28.7 (CH_2), 28.5 (CH_2), 25.7 (CH_2); ***m/z*** (ESI+, MeCN/H_2O) 387 ($[M+H]^+$, 25%), 287 ($[M-C_5H_9O_2]^+$, 100), 270 ($[M-C_5H_{10}O_3]^+$, 25).

^1H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[174]

4-{[(*tert*-Butoxy)carbonyl]amino}benzoic acid, **30**

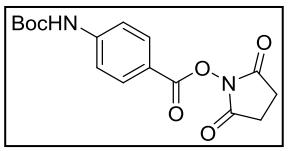


4-Aminobenzoic acid (3.43 g, 25.0 mmol) was dissolved in dioxane (50 mL) and water (25 mL). Triethylamine (7.0 mL, 50.2 mmol) was added to the solution, followed by di-*tert*-butyl dicarbonate (10.9 g, 50.0 mmol) and the reaction mixture was stirred at room temperature for 18 h. Solvent was removed *in vacuo*, and HCl (15 mL; 3 N, aq.) was added dropwise to the remaining residue. The resulting precipitate was filtered, washed with water (2 × 20 mL), hexane (2 × 20 mL), and finally dried *in vacuo* to afford the Boc-protected amine **30** as a colourless solid (5.86 g, 99%).

R_f (DCM:MeOH, 80:20) = 0.23; **mp** 187-189 °C, lit. [175] 191-192 °C; **IR** (neat, cm^{-1}) 3368 (NH), 2980-2497 (OH), 1705 (C=O), 1676 (C=O), 1609 (C=C), 1589 (C=C), 1524 (C=C); **^{1}H NMR** δ (500 MHz DMSO-d_6) 12.60 (1H, br s, CO_2H), 9.73 (1 H, s, N*H*), 7.84 (2H, d, *J* = 8.4 Hz, Ar*H*), 7.56 (2H, d, *J* = 8.4 Hz, Ar*H*), 1.50 (9H, s, $C(CH_3)_3$); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 167.5 (CO), 153.0 (CO), 144.2 (ArC), 130.8 (2ArCH), 124.5 (ArC), 117.7 (2ArCH), 80.1 (C), 28.5 ($3CH_3$); ***m/z*** (ESI-, MeCN/H_2O) 236 ($[M-H]^-$, 100%).

^1H and ^{13}C NMR spectroscopic data are in good agreement with the literature.[175]

2,5-Dioxopyrrolidin-1-yl 4-{[(*tert*-butoxy)carbonyl]amino}benzoate, **31**



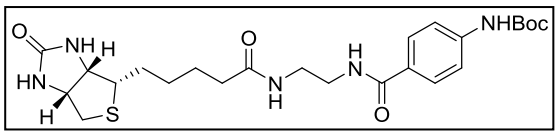
Carboxylic acid **30** (2.85 g, 12.0 mmol) and *N*-hydroxysuccinimide (1.38 g, 12.0 mmol) were dissolved in DCM (50 mL). DCC (2.48 g, 12.0 mmol) was added to the solution and the reaction mixture was stirred at room temperature for 18 h. The urea by-product was removed by filtration and the filtrate was concentrated to dryness *in vacuo* to give the NHS ester **31** as a colourless solid (3.13 g, 78%).

R_f (DCM:MeOH, 90:10) = 0.84; **mp** 172-174 °C, lit.[176] 172-173 °C; **IR** (neat, cm^{-1}) 3290 (NH), 1767 (C=O), 1728 (C=O), 1697 (C=O); **^{1}H NMR** δ (500 MHz, $CDCl_3$) 8.09 (2H, d, *J* = 8.9 Hz, Ar*H*), 7.53 (2H, d, *J* = 8.9 Hz, Ar*H*), 6.80 (1H, s, N*H*), 2.93 (4H, br s, $2CH_2$), 1.56 (9H, s, $C(CH_3)_3$); **^{13}C NMR** δ (126 MHz, $CDCl_3$) 169.4 (2CO), 161.4 (CO), 151.9 (CO), 144.6

(ArC), 132.2 (2ArCH), 118.8 (ArC), 117.5 (2ArCH), 81.7 (C), 28.3 ($3CH_3$), 25.7 ($2CH_2$); ***m/z*** (ESI+, MeCN/H_2O) 357 ($[M+Na]^+$, 100%).

^1H and ^{13}C NMR spectroscopic data is in good agreement with the literature.[176]

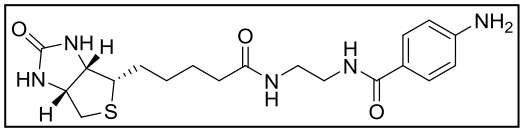
tert-Butyl *N*-{4-[(2-{5-[(3a*S*,4*S*,6a*R*)-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]pentanamido}ethyl) carbamoyl]phenyl}carbamate, **32**



Boc-protected amine **29** (1.87 g, 4.84 mmol) was suspended in DCM (50 mL) and TFA (5.6 mL, 73.1 mmol) was added. The solution was stirred at room temperature for 18 h and then concentrated *in vacuo*. The residue was dissolved in $NaHCO_3$ (65 mL; sat. aq.) and the NHS ester **31** (3.23 g, 9.67 mmol) was added in MeCN (60 mL). The reaction was monitored by TLC until judged to have gone to completion (~18 h). MeCN was removed *in vacuo* and the product was extracted in chloroform (3 × 60 mL), dried, and concentrated to dryness. Purification was performed by column chromatography (dry-loaded, DCM:MeOH, 90:10) to give the Boc-protected aniline **32** as a colourless solid (1.67 g, 68%).

R_f (DCM:MeOH, 90:10) = 0.42; **mp** 257-260 °C (decomp.); **$[\alpha]_D$** = +36.0 (c 1.00, DMF); **IR** (neat, cm^{-1}) 3327 (NH), 1692 (C=O), 1649 (C=O), 1632 (C=O), 1612 (C=C); **^{1}H NMR** (500 MHz, DMSO-d_6) δ 9.60 (1H, s, N*H*Boc), 8.31 (1H, t, *J* = 5.5 Hz, N*H*CO), 7.90 (1H, t, *J* = 5.7 Hz, CON*H*), 7.75 (2H, d, *J* = 8.7 Hz, 2Ar*H*), 7.51 (2H, d, *J* = 8.7 Hz, 2Ar*H*), 6.40 (1H, s, Biotin-N*H*), 6.35 (1H, s, Biotin-N*H*), 4.30 (1H, dd, *J* = 7.7, 5.1 Hz, C*H*CH_2S), 4.13-4.06 (1H, m, C*H*CHS), 3.31-3.25 (2H, m, CH_2NHCO), 3.24-3.19 (2H, m, CONHCH_2), 3.08-2.99 (1H, m, C*H*S), 2.81 (1H, dd, *J* = 12.4, 5.1 Hz, C$H_A$$H_B$S), 2.59 (1H, d, *J* = 12.4 Hz, C$H_A$$H_B$S), 2.14-2.02 (2H, m, C$H_2$CO), 1.64-1.39 (4H, m, 2CH_2), 1.49 (9H, s, C(CH_3)$_3$), 1.39-1.20 (2H, m, CH_2); **^{13}C NMR** (126 MHz, DMSO-d_6) 172.9 (CO), 166.4 (CO), 153.1 (CO), 142.7 (CO), 128.5 (2ArCH), 128.3 (ArC), 117.5 (2ArCH), 116.4 (ArC), 80.0 (C), 61.5 (CH), 59.7 (CH), 55.8 (CH), 40.3 (CH_2), 39.7 (CH_2), 38.8 (CH_2), 35.7 (CH_2), 28.6 (CH_2), 28.6 ($3CH_3$), 28.5 (CH_2), 25.7 (CH_2); **HRMS** (ESI+, MeOH) $[M+H]^+$ found 506.2437, $C_{24}H_{36}N_5O_5S$ requires 506.2432.

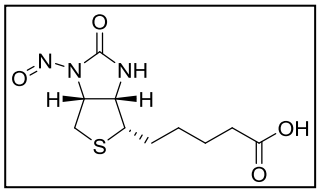
5-[(3a*S*,4*S*,6a*R*)-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazolidin-4-yl]-*N*-{2-[(4-aminophenyl)formamido] ethyl}pentanamide, **33**



Boc-protected aniline **32** (70 mg, 0.14 mmol) was suspended in DCM (2 mL). TFA (0.11 mL, 1.4 mmol) was added and the solution was stirred at room temperature for 18 h. The reaction mixture was cooled in an ice bath and excess acid was quenched by dropwise addition of Na_2CO_3 (10 mL; sat. aq.). The product was extracted in DCM and dried over $MgSO_4$. The mixture was filtered and solvent was removed from the eluate *in vacuo* to give the amine **33** as a colourless solid (48 mg, 85%).

R_f (DCM:MeOH, 80:20) = 0.47; **mp** 230-235 °C (decomp.); **$[\alpha]_D$ =** +30.0 (c 0.20, DMSO); **IR** (neat, cm^{-1}) 3246 (NH), 1628 (C=C); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 7.99 (1H, t, *J* = 5.5 Hz, N*H*CO), 7.89 (1H, t, *J* = 5.6 Hz, CON*H*), 7.55 (2H, d, *J* = 8.6 Hz, 2Ar*H*), 6.54 (2H, d, *J* = 8.6 Hz, 2Ar*H*), 6.42 (1H, s, Biotin-N*H*), 6.35 (1H, s, Biotin-N*H*), 5.59 (2H, s, NH_2), 4.33-4.27 (1H, m, C*H*CH$_2$S), 4.14-4.07 (1H, m, C*H*CHS), 3.29-3.22 (2H, m, CH_2), 3.22-3.14 (2H, m, CH_2), 3.09-3.01 (1H, m, C*H*S), 2.81 (1H, dd, *J* = 12.4, 5.1 Hz, CH_AH$_B$S), 2.58 (1H, m, CH$_A$$H_B$*S*), 2.07 (2H, t, *J* = 7.4 Hz, CH_2CO), 1.67-1.39 (3H, m, CH_2, CH_XH$_Y$), 1.39-1.18 (3H, m, CH_2, CH$_X$$H_Y$); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 172.8 (*C*O), 166.9 (*C*O), 163.2 (*C*O), 152.0 (Ar*C*), 129.1 (2Ar*C*H), 121.6 (Ar*C*), 113.0 (2Ar*C*H), 61.5 (*C*H), 59.7 (*C*H), 55.8 (*C*H), 40.3 (*C*H$_2$), 39.6 (*C*H$_2$), 39.0 (*C*H$_2$), 35.7 (*C*H$_2$), 28.6 (*C*H$_2$), 28.5 (*C*H$_2$), 25.7 (*C*H$_2$); **HRMS** (ESI+, MeOH) $[M+H]^+$ found 406.1905, $C_{19}H_{28}N_5O_3S$ requires 406.1907.

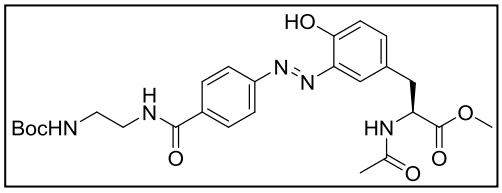
5-[(3a*S*,4*S*,6a*R*)-1-nitroso-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazol-4-yl]pentanoic acid, **34**



D-Biotin (50 mg, 0.205 mmol) was suspended in water (4 mL) and the solution was cooled to 0 °C. TFA (40 µL, 0.520 mmol) was added, followed by $NaNO_2$ (25 mg, 0.350 mmol) and the reaction mixture was stirred at 0 °C for 2 h, monitored by TLC. The product precipitated from solution, and was isolated by filtration and washed with diethyl ether. The solid was dried *in vacuo* to give nitrosobiotin **34** as a pale yellow solid (51 mg, 91%).

R_f (DCM:MeOH:AcOH, 98:2:2) = 0.35; **IR** (neat, cm^{-1}) 3550 (NH), 3410 (NH), 3298 (OH), 1759 (C=O), 1717 (C=N), 1379 (N=O), 1341 (N=O); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 11.98 (1H, s, CO_2H), 8.80 (1H, s, N*H*), 4.83 (1H, ddd, *J* = 7.0, 5.9, 0.9 Hz, C*H*CH_2S), 4.27 (1H, ddd, *J* = 7.0, 4.3, 1.4 Hz, C*H*CHS), 3.23 (1H, ddd, *J* = 8.9, 6.0, 4.3 Hz, C*H*S), 2.98 (1H, dd, *J* = 13.3, 5.9 Hz, C$H_A$$H_B$S), 2.68 (1H, d, *J* = 13.3 Hz, C$H_A$$H_B$), 2.23 (2H, t, *J* = 7.4 Hz, C$H_2$$CO_2$H), 1.77-1.66 (1H, m, SCHC$H_A$$H_B$), 1.58-1.46 (3H, m, SCHC$H_A$$H_B$, C$H_2$$CH_2CO_2$H), 1.44-1.28 (2H, m, C$H_2$$CH_2CH_2CO_2$H); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 174.9 (*C*O), 154.2 (*C*O), 60.0 (*C*HCH_2S), 59.0 (*C*HCHS), 54.4 (*C*HS), 35.0 (*C*H_2S), 33.9 (*C*H_2), 28.6 (*C*H_2), 28.1 (*C*H_2), 24.9 (*C*H_2); **HRMS** (ESI−, MeCN/H_2O) $[M-H]^-$ found 272.0689, $C_{10}H_{14}N_3O_4S$ requires 272.0700.

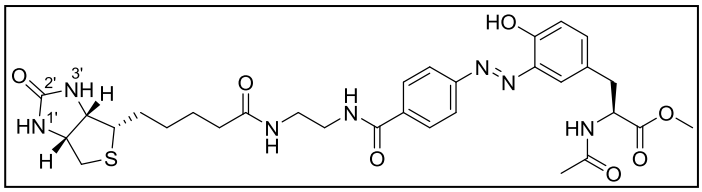
Methyl (2*S*)-3-[3-(2-{4-[(2-{[(*tert*-butoxy)carbonyl]amino}ethyl)carbamoyl]phenyl}diazen-1-yl)-4-hydroxyphenyl]-2-acetamidopropanoate, **35**



Carboxylic acid **7** (50 mg, 0.130 mmol) was dissolved in DMF (1 mL), and EDC (35 mg, 0.183 mmol) and HOBt (10 mg, 0.074 mmol) were added. The Boc-protected diamine **28** (52 mg, 0.324 mmol) was dissolved in DMF (1 mL) and added to the reaction mixture. The reaction was monitored by TLC, and the carboxylic acid was consumed after 48 h. Solvent was removed *in vacuo* and the product was purified by preparative RP-HPLC (System 4) using an isocratic solvent system (H_2O + 0.1%: MeCN + 0.1% TFA, 57:43). The collected fractions were lyophilised to give the amide coupling product **35** as an orange amorphous solid (33 mg, 48%).

R_f (DCM:MeOH, 95:5) = 0.54; **[α]$_D$ =** +37.5 (c 0.40, DMSO); **IR** (neat, cm^{-1}) ; **^{1}H NMR** δ (500 MHz, $CDCl_3$) 12.65 (1H, s, O*H*), 7.99 (2H, d, *J* = 8.3 Hz, 2Ar*H*), 7.89 (2H, d, *J* = 8.3 Hz, 2Ar*H*), 7.72 (1H, d, *J* = 2.3 Hz, Ar*H*), 7.56 (1H, br s, BocN*H*), 7.14 (1H, dd, *J* = 8.5, 2.3 Hz, Ar*H*), 6.97 (1H, d, *J* = 8.5 Hz, Ar*H*), 6.19 (1H, d, *J* = 7.7 Hz, AcN*H*), 5.17 (1H, t, *J* = 6.2 Hz, N*H*COAr), 4.94 (1H, dt, *J* = 7.7, 5.8 Hz, Tyr(α-C*H*)), 3.78 (3H, s, OCH_3), 3.65 – 3.57 (2H, m, BocNHCH_2), 3.49 – 3.42 (2H, m, CH_2NHCOAr), 3.22 (1H, dd, *J* = 14.1, 5.8 Hz, Tyr(β-CH_AH$_B$)), 3.15 (1H, dd, *J* = 14.1, 5.8 Hz, Tyr(β-CH$_A$$H_B$), 2.04 (3H, s, C$H_3$CONH), 1.46 (9H, s, Boc(CH_3)$_3$); **^{13}C NMR** δ (126 MHz, $CDCl_3$) 172.1 (*C*O), 169.8 (*C*O), 166.8 (*C*O), 157.8 (*C*O), 152.2 (Ar*C*), 151.9 (Ar*C*), 137.2 (Ar*C*), 136.2 (Ar*C*), 134.8 (Ar*C*H), 133.8 (Ar*C*H), 128.3 (2Ar*C*H), 127.7 (ArC), 122.3 (2Ar*C*H), 118.5 (Ar*C*H), 80.2 (*C*), 53.3 (*C*H), 52.5 (*C*H), 42.6 (*C*H_2), 39.9 (*C*H_2), 36.9 (*C*H_2), 28.4 (3*C*H_3), 23.2 (*C*H_3); **HRMS** (ESI+, MeCN/H_2O) $[M+Na]^+$ found 550.2275, $C_{26}H_{33}N_5O_7Na$ requires 550.2272.

Methyl (2*S*)-3-[3-(2-{4-[(2-{5-[(3a*S*,4*S*,6a*R*)-2-oxo-hexahydro-1*H*-thieno[3,4-*d*]imidazol-4-yl]pentanamido}ethyl)carbamoyl]phenyl}diazen-1-yl)-4-hydroxyphenyl]-2-acetamidopropanoate, **36**



Boc-protected azobenzene **35** (110 mg, 0.290 mmol) was dissolved in DCM (6 mL) and TFA (0.4 mL, 4.31 mmol) was added. The reaction was stirred at rt for 3 h and monitored by TLC with Ninhydrin staining. The solvent and TFA were removed under a stream of N_2, and the residue was redissolved in $NaHCO_3$ (8 mL; sat. aq.). Biotin-NHS **27** (187 mg, 0.55 mmol) was dissolved in DMF (2 mL) and added to the reaction mixture. The reaction mixture was stirred at room temperature for 18 h and monitored by TLC. The reaction mixture was acidified with HCl (1 M, aq.) and the product was extracted in EtOAc. Purification was performed by column chromatography (dry-loaded, DCM:MeOH, 95:5) to give the biotinylated azobenzene **36** as an orange solid (176 mg, 93%).

R_f (DCM:MeOH, 90:10) = 0.30; **IR** (neat, cm^{-1}) 3289 (OH), 1676 (CO), 1643 (CO), 1541 (C=C); **^{1}H NMR** δ (500 MHz, DMSO-d_6) 10.92 (1H, s, O*H*), 8.66 (1H, t, *J* = 5.6 Hz, N*H*COAr), 8.36 (1H, d, *J* = 7.8 Hz, AcN*H*), 8.09 – 7.99 (4H, m, 4Ar*H*), 7.95 (1H, t, *J* = 5.8 Hz, CON*H*CH$_2$), 7.61 (1H, d, *J* = 2.3 Hz, Ar*H*), 7.32 (1H, dd, *J* = 8.5, 2.3 Hz, Ar*H*), 7.02 (1H, d, *J* = 8.5 Hz, Ar*H*), 6.41 (1H, s, Biotin-N(3')*H*), 6.34 (1H, s, Biotin-N(1')*H*), 4.47 (1H, ddd, *J* = 9.2, 7.8, 5.6 Hz, Tyr(α-C*H*), 4.29 (1H, dd, *J* = 7.8, 5.1 Hz, Biotin-C*H*CH$_2$S), 4.13-4.09

(1H, m, Biotin-C*H*CHS), 3.62 (3H, s, OCH_3), 3.39-3.33 (2H, m, CH_2NHCO), 3.29-3.22 (2H, m, CONHCH_2), 3.09-3.02 (1H, m, Biotin-C*H*S) 3.02 (1H, dd, *J* = 13.8, 5.6 Hz, Tyr(β-CH_AH$_B$)), 2.89 (1H, dd, *J* = 13.8, 9.2 Hz, Tyr(β-CH$_A$$H_B$)), 2.80 (1H, dd, *J* = 12.5, 5.1 Hz, Biotin-CH_AH$_B$S), 2.57 (1H, d, *J* = 12.5 Hz, Biotin-CH$_A$$H_B$S), 2.09 (2H, t, *J* = 7.4 Hz, CH_2CONH), 1.81 (3H, s, CH_3CONH), 1.66-1.57 (1H, m, SCHCH_AH$_B$), 1.57-1.41 (3H, m, SCHCH$_A$$H_B$, SCHCH$_2CH_2$), 1.39-1.20 (2H, m, C$H_2CH_2$CONH); **^{13}C NMR** δ (126 MHz, DMSO-d_6) 172.9 (*C*O), 172.6 (*C*O), 169.8 (*C*O), 166.1 (*C*O), 163.2 (*C*O), 154.3 (Ar*C*), 153.5 (Ar*C*), 138.7 (Ar*C*), 136.8 (Ar*C*), 135.4 (Ar*C*H), 129.2 (Ar*C*), 128.9 (2Ar*C*H), 122.9 (2Ar*C*H), 122.6 (Ar*C*H), 118.7 (Ar*C*H), 61.5 (*C*H), 59.7 (*C*H), 55.9 (*C*H), 54.1 (*C*H), 52.3 (*C*H$_3$), 40.3 (*C*H$_2$), 39.9 (*C*H$_2$), 38.6 (*C*H$_2$), 36.3 (*C*H$_2$), 35.7 (*C*H$_2$), 28.6 (*C*H$_2$), 28.5 (*C*H$_2$), 25.7 (*C*H$_2$), 22.7 (*C*H$_3$); **HRMS** (ESI+, MeCN/H$_2$O) $[M+Na]^+$ found 676.2543, $C_{31}H_{39}N_7O_7SNa$ requires 676.2524.

Ac-Tyr(3-NH$_2$)-OMe and Leu-Enk[Tyr(3-NH$_2$)] UHPLC standards

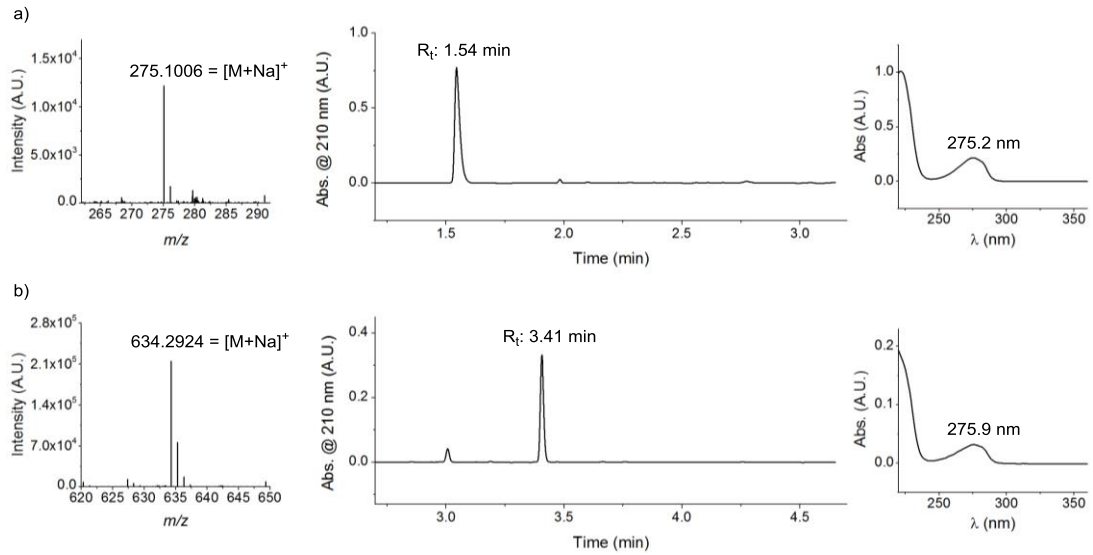


Figure 5-1 *UHPLC and MS of the standards used for SAv resin and Affi-Gel catch-and-release. (a) The 3-aminotyrosine-derived standard* ***5*** *gave a retention time of 1.54 min. (b) The 3-aminotyrosine modified Leu-Enkephalin* ***37*** *was obtained from a preformed biotin-Leu-Enkephalin azobenzene linker, which was incubated with SAv resin and cleaved in $Na_2S_2O_4$. The sample was desalted using a C_{18} flash cartridge, with elution in MeOH to give a retention time of 3.41 min with mass confirmation by ESI-MS.*

5.8 Experimental procedures for Chapter 4

5.8.1 Solvent-accessible surface area calculation

Using PyMOL software, the SASA calculation was performed using the automated SASA script. The desired .pdb file was retrieved from the RCSB Protein Data Bank, and saved in the same directory as the SASA script 'get_resn_sasa.py'. The .pdb file was opened in PyMOL and the script was called using the command 'run get_resn_sasa.py'. Upon loading the script, the protein backbone is shown as a cartoon, with colour-coded stick representations for Tyr (grey), His (purple), Lys (green) and Cys (yellow) residues, as well as the N-terminus (pink). The SASA result of the α-NH_2 of the N-terminus is displayed automatically in the External GUI, and SASA calculation of the Tyr residues can be performed by entering the command 'get_tyr_sasa'. Alternatively, 'get_his_sasa', 'get_lys_sasa' or 'get_cys_sasa' can be used to calculate the SASA of other residues. The SASA output is displayed in the GUI, and a .txt file of the results is also created in the same directory as the script and .pdb file.

Automated SASA script

The automated SASA script is given below, with annotation of the 'get_tyr_sasa' code in **bold**.

```
'''         #Multi-line string, instead of using hashtag

Authors:
            Miroslav Kosar, Christopher Allan
            The University of Edinburgh
Usage:
            This script calculates the solvent-accessible surface area of the
selected atoms.
            The output results include the SASA value of each selected atom,
the sum SASA value
            corresponding to the amino acid in question, and a sum of all
residues of that type
            over the whole protein sequence.
            Five function are defined:
            get_tyr_sasa    - calculates the SASA at the ortho positions of
the tyrosine
                              phenol side chain
            get_his_sasa    - calculates the SASA at the CE1-position of the
imidazole
                              side chain of histidine.
            get_lys_sasa     - calculates the SASA at the terminal amine of
the lysine side chain
            get_cys_sasa     - calculates the SASA at the terminal thiol of
the cysteine side chain
            get_nterm_sasa  - calculates the SASA at the N-terminus

'''

#Import Pymol module to pull out Pymol data and functions
from pymol import cmd
from pymol import stored
import time
start = time.time()

#Visuals
#'cmd.do' sends command to Pymol
cmd.do('hide all')
cmd.do('show cartoon')
cmd.do('colour gray90, all')
cmd.do('util.cbaw resn tyr')
```

```
cmd.do('util.cbab resn his')
cmd.do('util.cbag resn lys')
cmd.do('util.cbay resn cys')
cmd.do('select nterm, resi 1')
cmd.do('util.cbap nterm')
cmd.do('select none')
cmd.do('show sticks, resn tyr')
cmd.do('show sticks, resn his')
cmd.do('show sticks, resn lys')
cmd.do('show sticks, resn cys')
cmd.do('show sticks, nterm')

#def function get_tyr_sasa(resn="TYR", myname="CE1+CE2"):
def get_tyr_sasa(resn="TYR", myname="CE1+CE2"):
    '''(str, str)->Nonetype

    '''
    #set 'dot_solvent' to 0 for AREA or 1 for SASA
    #set 'solvent_radius' to 1.4 to model water
    #set 'dot_density' to 4 for maximum accuracy,
    #however this causes long calculation time
    cmd.set('dot_solvent', 1)
    cmd.set('solvent_radius', 1.4)
    cmd.set('dot_density', 2)

    #Open new .txt file in 'write' mode ('w')
    outfile='tyr_sasa_out.txt'
    f = open(outfile, 'w')
    #'target_sel' and 'expression' used to create output, '%s' refers to
    #item in Tuple (string)
    target_sel = 'resname %s and name %s' % (resn, myname)
    #Convert resn (e.g. Tyr) and atomname (e.g. CE1) to all uppercase
    #characters
    resn = resn.upper()
    atomname = myname.upper()
    #Create empty list called 'stored.infolist'
    stored.infolist = []
    #Add pdb ref. (model), chain, residue number (resv), amino acid (resn),
    #atom name (name) to list
    expression = "stored.infolist.append([model,chain,resv,resn,name])"
    #iterate for each atom, populates list
    cmd.iterate(target_sel, expression)

    #Store information as a dictionary, '{_' is empty dictionary, populate
    #with 'keys' and 'values'
    sasa_per_resi={_
    #Create list for keys, fill with strings from 'stored.infolist'
    #(residue[1] = chain, residue[3] = resn, residue[2] = resv)
    #'%4s'is a string from a Tuple, formatted to 4 characters (e.g.0001)
    keys = []
    for residue in stored.infolist:
        key ='Chain-%s-%s-%4s' % (residue [1], residue[3], residue[2])
        #Safety artificial values set to keys to check for malfunction and
        #Zero SASA prediction
        if key not in sasa_per_resi:
            sasa_per_resi[key]={_
            sasa_per_resi[key]['CE1']=-999.00
            sasa_per_resi[key]['CE2']=-999.00
            sasa_per_resi[key]['SUM']=-999.00
            keys.append(key)
        #'i' is generic name for variable, residue[2] = resv,
        #residue[4] = atom name
        i="resi %s and name %s" % (residue[2], residue[4])
        #Overwrite artificial values with actual SASA prediction
        #'cmd.get_area' is SASA command, '(i)' holds (resi, atom name)
        value = cmd.get_area(i)
        #Combine atom ID with SASA value in dictionary
        sasa_per_resi[key][residue[4]] = value

    #Accessing values and formatting data output
    for key in keys:
        #Calculate sum SASA by adding SASA of each ortho position (CE1 + CE2)
        sasa_sum = sasa_per_resi[key]['CE1'] + sasa_per_resi[key]['CE2']
        sasa_per_resi[key]['SUM'] = sasa_sum
```

```
        #Output format, '%s' = string from Tuple, '%f' = floating point value
        #from Tuple, '5.2' refers to formatting, 5 characters, 2 of which are
        #decimal places (e.g.000.01)
        line = '%s -> CE1: %5.2f    CE2: %5.2f    SUM: %5.2f' % (key,
sasa_per_resi[key]['CE1'],

sasa_per_resi[key]['CE2'],

sasa_per_resi[key]['SUM'])
        #Print output ('line') and new line ('\n')
        print line
        f.write(line +'\n')

    #Get the total SASA for resn of the protein
    total_sasa = 0.00
    for key in sasa_per_resi:
        for key in keys:
            total_sasa  =  total_sasa  +  sasa_per_resi[key]['CE1']  +
sasa_per_resi[key]['CE2']
        break

    out = 'Total %s SASA: %5.2f' % (resn, total_sasa)
    print out
    #Get total number of tyr residues (/2 as Tyr has 2 reactive position,
    #CE1+CE2, not required for other reactive residues
    total_tyr = 'Total %s residues: %d' % (resn, len(stored.infolist)/2)
    print total_tyr

    f.write(out + '\n')
    f.write(total_tyr + '\n')
    f.write('total execution time %s seconds'  %(time.time()-start))
    f.close()

#Bind user-defined function as a command (string name, function)
cmd.extend('get_tyr_sasa', get_tyr_sasa)


def get_his_sasa(resn="HIS", myname="CE1"):
    '''(str, str)->Nonetype

    '''
    cmd.set('dot_solvent', 1)
    cmd.set('dot_density', 2)

    outfile='his_sasa_out.txt'
    f = open(outfile, 'w')
    target_sel = 'resname %s and name %s' % (resn, myname)
    resn = resn.upper()
    atomname = myname.upper()
    stored.infolist = []
    expression = "stored.infolist.append([model,chain,resv,resn,name])"
    cmd.iterate(target_sel, expression)

    sasa_per_resi={_
    keys = []
    for residue in stored.infolist:
        key ='Chain-%s-%s-%4s' % (residue[1], residue[3], residue[2])
        if key not in sasa_per_resi:
            sasa_per_resi[key]={_
            sasa_per_resi[key]['CE1']=-999.00
            keys.append(key)
        i="resi %s and name %s" % (residue[2], residue[4])
        value = cmd.get_area(i)
        sasa_per_resi[key][residue[4]] = value

    for key in keys:
        line = '%s -> CE1: %5.2f' % (key, sasa_per_resi[key]['CE1'])

        print line
        f.write(line +'\n')

    total_sasa = 0.00
    for key in sasa_per_resi:
        for key in keys:
```

```
            total_sasa = total_sasa + sasa_per_resi[key]['CE1']
        break

    out = 'Total %s SASA: %f' % (resn, total_sasa)
    print out
    total_his = 'Total %s residues: %d' % (resn, len(stored.infolist))
    print total_his

    f.write(total_his + '\n')
    f.write(out + '\n')
    f.close()

cmd.extend('get_his_sasa', get_his_sasa)


def get_lys_sasa(resn="LYS", myname="NZ"):
    '''(str, str)->Nonetype

    '''
    cmd.set('dot_solvent', 1)
    cmd.set('dot_density', 2)

    outfile='lys_sasa_out.txt'
    f = open(outfile, 'w')
    target_sel = 'resname %s and name %s' % (resn, myname)
    resn = resn.upper()
    atomname = myname.upper()
    stored.infolist = []
    expression = "stored.infolist.append([model,chain,resv,resn,name])"
    cmd.iterate(target_sel, expression)

    sasa_per_resi={_
    keys = []
    for residue in stored.infolist:
        key ='Chain-%s-%s-%4s' % (residue[1], residue[3], residue[2])
        if key not in sasa_per_resi:
            sasa_per_resi[key]={_
            sasa_per_resi[key]['NZ']=-999.00
            keys.append(key)
        i="resi %s and name %s" % (residue[2], residue[4])
        value = cmd.get_area(i)
        sasa_per_resi[key][residue[4]] = value

    for key in keys:
        line = '%s -> NZ: %5.2f' % (key, sasa_per_resi[key]['NZ'])

        print line
        f.write(line +'\n')

    total_sasa = 0.00
    for key in sasa_per_resi:
        for key in keys:
            total_sasa = total_sasa + sasa_per_resi[key]['NZ']
        break

    out = 'Total %s SASA: %f' % (resn, total_sasa)
    print out
    total_lys = 'Total %s residues: %d' % (resn, len(stored.infolist))
    print total_lys

    f.write(total_lys + '\n')
    f.write(out + '\n')
    f.close()

cmd.extend('get_lys_sasa', get_lys_sasa)


def get_cys_sasa(resn="CYS", myname="SG"):
    '''(str, str)->Nonetype

    '''
    cmd.set('dot_solvent', 1)
    cmd.set('dot_density', 2)
```

```
    outfile='cys_sasa_out.txt'
    f = open(outfile, 'w')
    target_sel = 'resname %s and name %s' % (resn, myname)
    resn = resn.upper()
    atomname = myname.upper()
    stored.infolist = []
    expression = "stored.infolist.append([model,chain,resv,resn,name])"
    cmd.iterate(target_sel, expression)

    sasa_per_resi={_
    keys = []
    for residue in stored.infolist:
        key ='Chain-%s-%s-%4s' % (residue[1], residue[3], residue[2])
        if key not in sasa_per_resi:
            sasa_per_resi[key]={_
            sasa_per_resi[key]['SG']=-999.00
            keys.append(key)
        i="resi %s and name %s" % (residue[2], residue[4])
        value = cmd.get_area(i)
        sasa_per_resi[key][residue[4]] = value

    for key in keys:
        line = '%s -> SG: %5.2f' % (key, sasa_per_resi[key]['SG'])

        print line
        f.write(line +'\n')

    total_sasa = 0.00
    for key in sasa_per_resi:
        for key in keys:
            total_sasa = total_sasa + sasa_per_resi[key]['SG']
        break

    out = 'Total %s SASA: %f' % (resn, total_sasa)
    print out
    total_cys = 'Total %s residues: %d' % (resn, len(stored.infolist))
    print total_cys

    f.write(total_cys + '\n')
    f.write(out + '\n')
    f.close()

cmd.extend('get_cys_sasa', get_cys_sasa)


def get_nterm_sasa(resi="1", name="N"):
    '''(str, str)->Nonetype

    '''
cmd.set('dot_solvent', 1)
cmd.set('dot_density', 2)

cmd.select("nterm", "resi 1 and name n")
nterm_area=cmd.get_area("nterm")
print 'N-terminus SASA', nterm_area

cmd.extend('get_nterm_sasa', get_nterm_sasa)
```

Tyr SASA results for protein substrates

Ribonuclease A (PDB: 4J5Z)

Residue	SASA (Å^2)
Y025	3.02
Y073	7.76
Y076	29.36
Y092	24.98
Y097	0.96
Y115	21.68

Soybean trypsin inhibitor (PDB: 1BA7)

Residue	SASA (Å^2)
Y017	0.00
Y018	7.23
Y062	38.09
Y131	0.00

β-Lactoglobulin A (PDB: 3NPO)

Residue	SASA (Å^2)
Y020	6.11
Y042	0.00
Y099	15.33
Y102	0.00

Myoglobin (PDB: 5D5R)

Residue	SASA (Å^2)
Y103	8.21
Y146	0.00

Hen egg white lysozyme (PDB: 4YM8)

Residue	SASA (Å^2)
Y020	21.72
Y023	13.44
Y053	13.55

Cytochome C (PDB: 2B4Z)

Residue	SASA (Å^2)
Y048	4.59
Y067	11.62
Y074	8.31
Y097	9.98

5.8.2 TNM modification and reduction of RNase A

The TNM modification and reduction protocol was adapted from Francis *et al*.[122] TNM (10 µL, 3.5% v/v in H_2O) was added to a solution of RNase A (570 µL, 255 µM RNase A in 100 mM phosphate buffer, pH 8.8) and the reaction mixture was agitated on a rotator for 24 h at rt. Excess TNM was removed by successive concentration and washing using a MWCO spin cartridge, phosphate buffer (3 × 400 µL, 100 mM, pH 7.2) and water (2 x 400 µL). The nitrated protein was lyophilised and re-dissolved in aqueous $Na_2S_2O_4$ (300 µL, 100 mM), briefly mixed by vortex and agitated on a rotator for 2 h at rt. The sample volume was reduced to ~50 µL on a MWCO spin cartridge and desalted by washing with phosphate buffer (5 x 400 µL, 100 mM, pH 7.2).

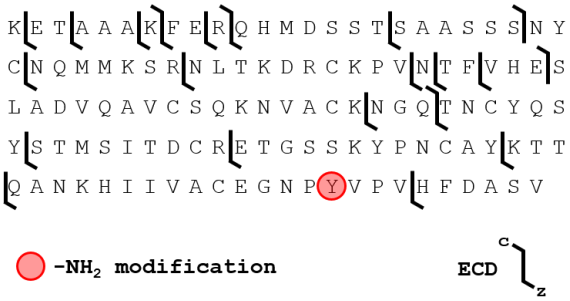
5.8.3 Top-down fragmentation

A sample of denatured and reduced catch-and-release modified RNase A was directly infused into a Bruker SolariX 12T FT-ICR mass spectrometer using a Triversa Nanomate (Advion Biosciences) and an isolated species was fragmented by ECD and CID.

ECD fragmentation

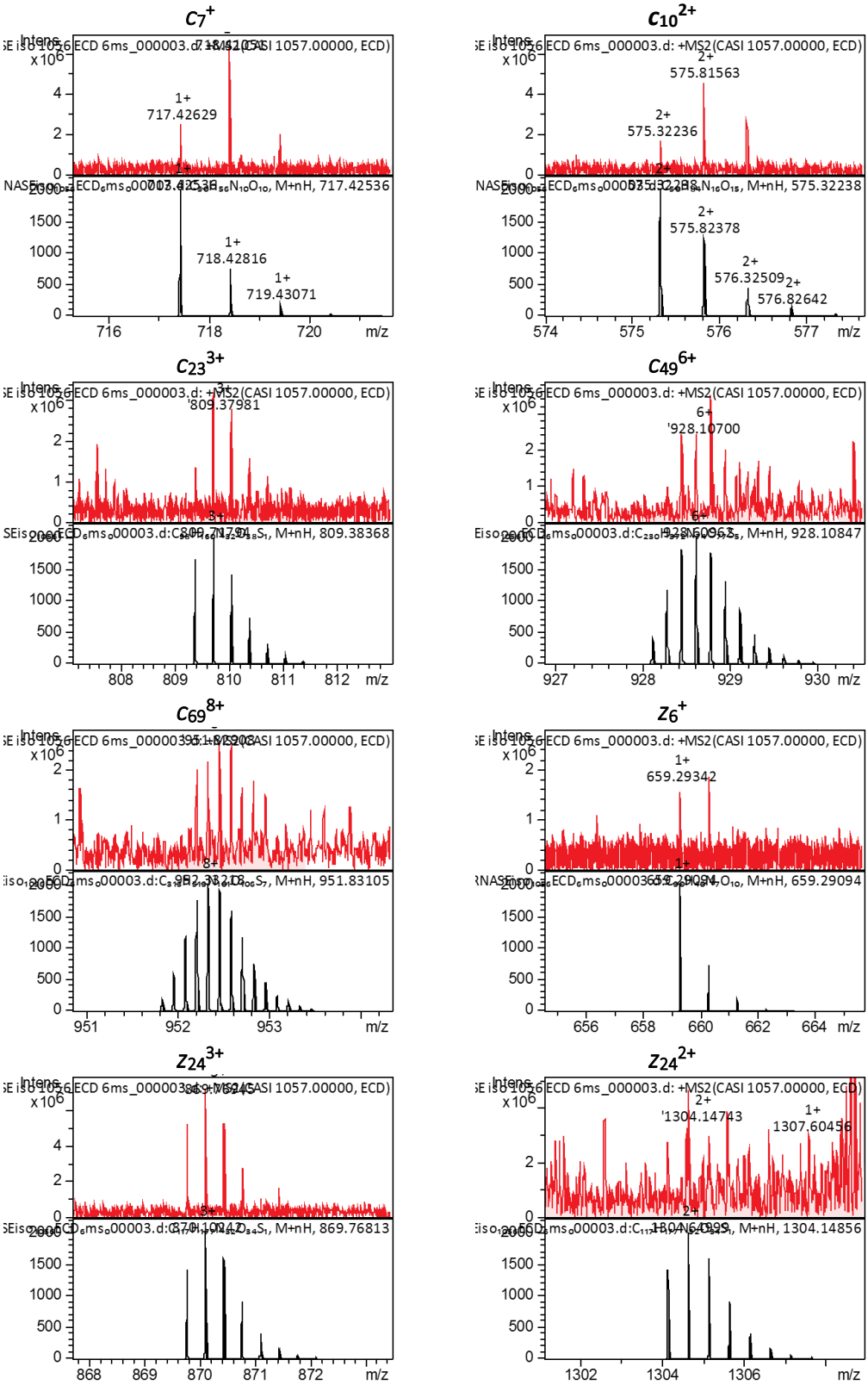
The 13+ charge state of the singly modified protein was isolated with ion accumulation time of up to 5 sec. 1.7 A was applied to the cathode, the bias voltage was set to 1 V, and the lens voltage was 20 V. Dissociation of the isolated species was achieved by applying a pulse for 6 ms, ECD time was 8 ms. The resulting peptide fragments were calibrated by single point calibration to $[C_{131}H_{206}N_{36}O_{39}S_1]^{3+}$: 979.831572 *m/z*. Using DataAnalysis software (version 4.1, Bruker Daltronics), a list of monoisotopic masses was generated and this was searched against the theoretical masses of all expected peptides using ProSight PTM software (Kelleher Lab, Northwestern University)[168, 169] with a mass error tolerance of 5 ppm. The identified peptide peaks were further scrutinised by comparison with the theoretical isotope distribution using DataAnalysis.

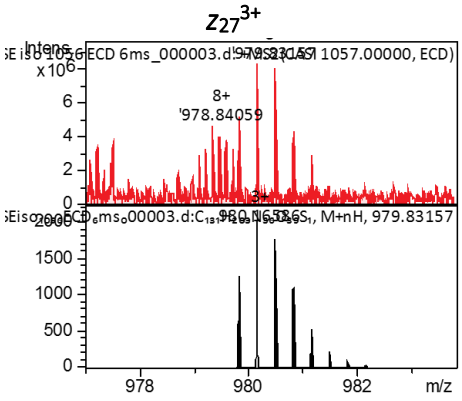
ECD: Identified peptide ions

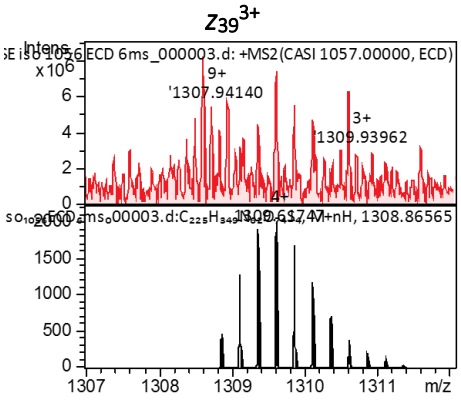


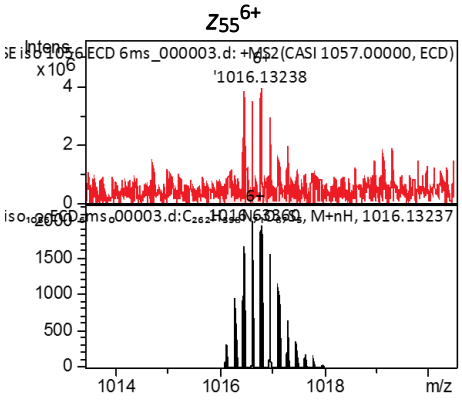
Peptide ions identified by top-down fragmentation of singly modified RNase A by ECD.

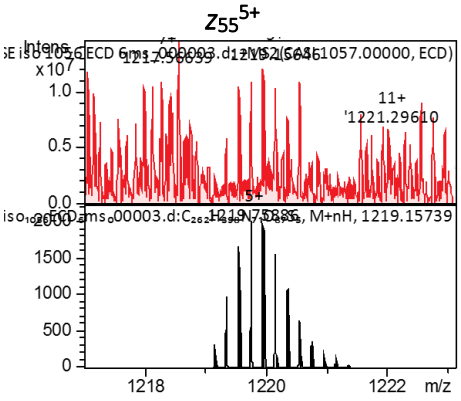
Peptide Ion	*m/z*	*z*	Measured Molecular Mass (Da)	Theoretical Molecular Mass (Da)	Mass Error (ppm)
c7	717.4263	1+	716.4193	716.4179	1.98
c10	575.3224	2+	1148.6305	1148.6300	0.41
c23	809.3798	3+	2425.1179	2425.1290	-4.58
c49	928.1070	6+	5562.5986	5562.6069	-1.50
c69	951.8291	8+	7606.5747	7606.5899	-2.00
z6	659.2934	1+	658.2864	658.2837	4.21
z24	869.7695	3+	2606.2847	2606.2816	1.19
	1304.1474	2+	2606.2847	2606.2816	1.19
z27	979.8316	3+	2936.4731	2936.4719	0.41
z39	1413.3478	3+	4237.0218	4237.0113	2.46
z48	1308.8655	4+	5231.4332	5231.4326	0.12
z55	1016.1324	6+	6090.7481	6090.7496	-0.25
	1219.1565	5+	6090.7481	6090.7496	-0.25
	1523.6951	4+	6090.7481	6090.7496	-0.25
z58	1278.9818	5+	6389.8690	6389.8726	-0.57
	1598.4726	4+	6389.8690	6389.8726	-0.57
z78	1417.6533	6+	8499.8766	8499.9025	-3.05
z80	1459.0057	6+	8747.9910	8748.0186	-3.16
z81	1267.0153	7+	8862.0552	8862.0615	-0.71
z91	1431.9588	7+	10016.6608	10016.6845	-2.37
z98	1211.2408	9+	10892.1019	10892.0951	0.62
z107	1177.2463	10+	11762.3990	11762.4129	-1.19
z114	1255.8772	10+	12548.6967	12548.7096	-1.03
z115	1155.9895	11+	12704.8047	12704.8107	-0.47
z117	1181.0907	11+	12980.9114	12980.9217	-0.79
z118	1192.7363	11+	13109.0107	13109.0166	-0.45
z121	1212.1096	11+	13322.1257	13322.1280	-0.17
	1333.2197	10+	13322.1257	13322.1280	-0.17
z123	1233.0243	11+	13552.1880	13552.2182	-2.23

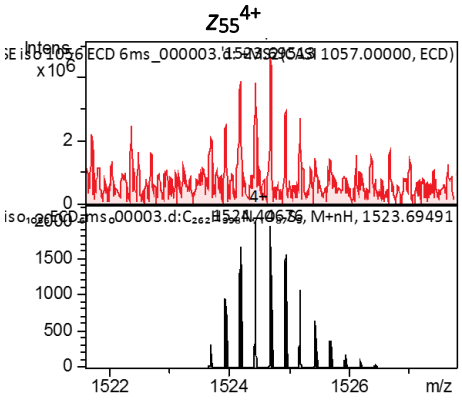

c_7^{+}
c_{10}^{2+}
c_{23}^{3+}
c_{49}^{6+}
c_{69}^{8+}
z_6^{+}
z_{24}^{3+}
z_{24}^{2+}
1+
717.42629
718.42816
719.43071
2+
575.81563
575.32236
575.82378
576.32509
576.82642
3+
809.37981
6+
928.10700
8+
659.29342
1304.14743
1307.60456
m/z

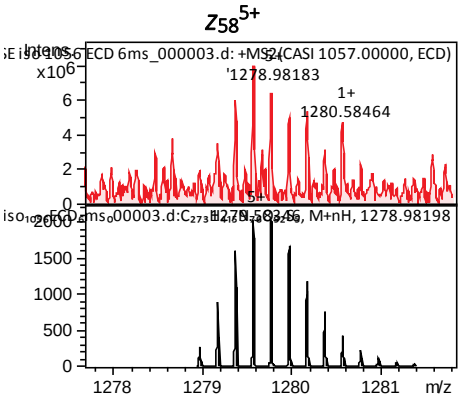

z_{27}^{3+}
8+
'978.84059
3+
979.83157, M+nH, 979.83157
978
980
982
m/z

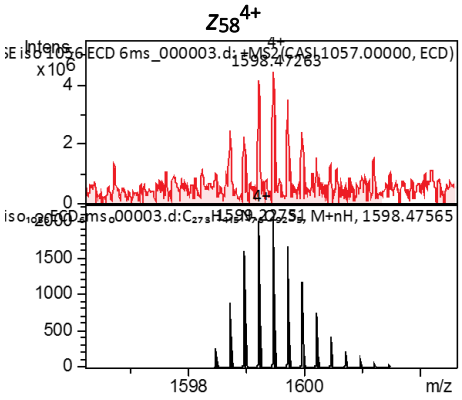

z_{39}^{3+}
9+
'1307.94140
3+
'1309.93962
4+
1308.86565, M+nH, 1308.86565
1307
1308
1309
1310
1311
m/z

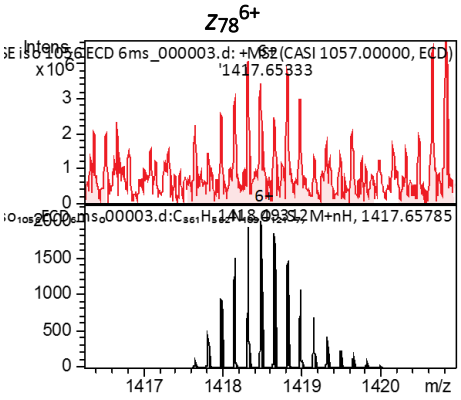

z_{55}^{6+}
6+
'1016.13238
6+
1016.13237, M+nH, 1016.13237
1014
1016
1018
m/z

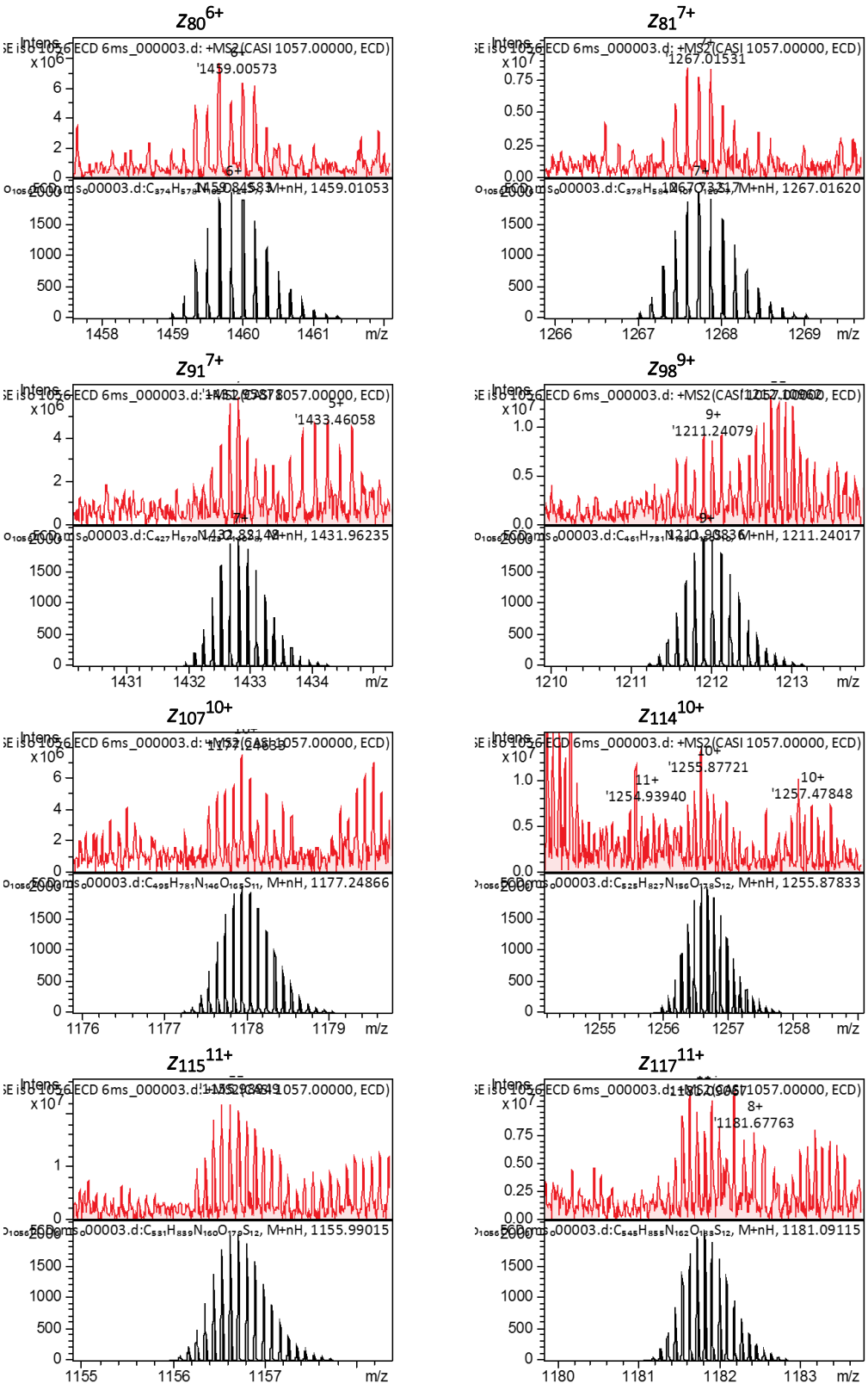

z_{55}^{5+}
11+
'1221.29610
5+
1219.15739, M+nH, 1219.15739
1218
1220
1222
m/z

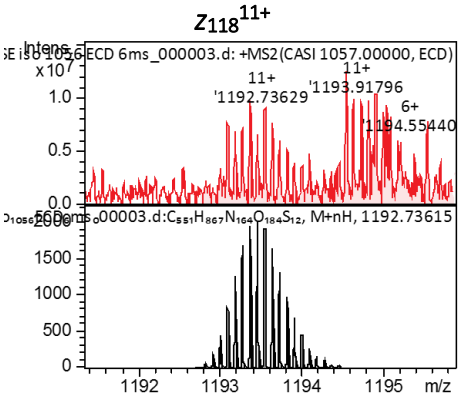

z_{55}^{4+}
4+
1523.69491, M+nH, 1523.69491
1522
1524
1526
m/z

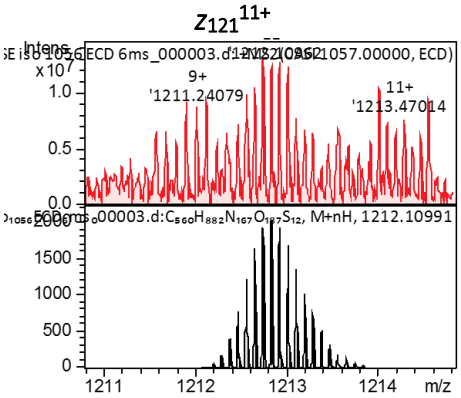

z_{58}^{5+}
'1278.98183
1+
1280.58464
5+
1278.98198, M+nH, 1278.98198
1278
1279
1280
1281
m/z

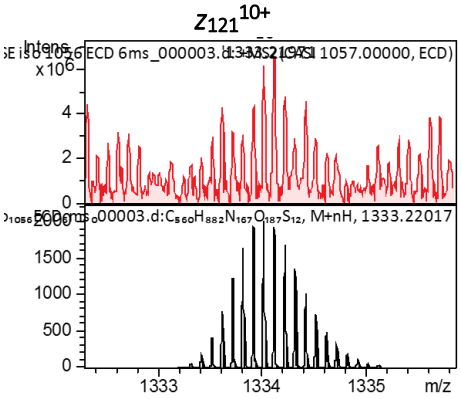

z_{58}^{4+}
4+
1598.47263
4+
1598.47565, M+nH, 1598.47565
1598
1600
m/z

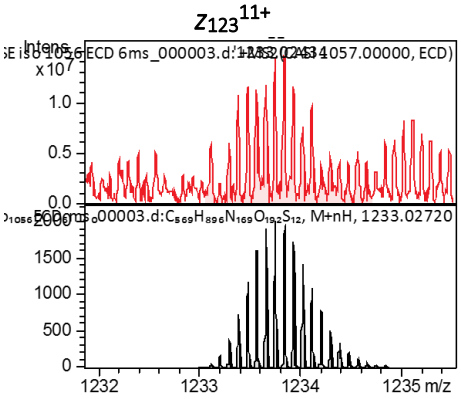

z_{78}^{6+}
'1417.65333
6+
1417.65785, M+nH, 1417.65785
1417
1418
1419
1420
m/z


z_{80}^{6+}
'1459.00573
6+
z_{81}^{7+}
'1267.01531
7+
z_{91}^{7+}
'1433.46058
7+
z_{98}^{9+}
'1211.24079
9+
z_{107}^{10+}
z_{114}^{10+}
'1255.87721
'1254.93940
'1257.47848
z_{115}^{11+}
z_{117}^{11+}
'1181.67763
8+
m/z


z_{118}^{11+}
Intens.
x10^7
1.0
0.5
0.0
11+
'1192.73629
11+
'1193.91796
6+
'1194.55440
00003.d:C551H867N164O184S12, M+nH, 1192.73615
2000
1500
1000
500
0
1192
1193
1194
1195
m/z


z_{121}^{11+}
Intens.
x10^7
1.0
0.5
0.0
9+
'1211.24079
11+
'1213.47014
00003.d:C560H882N167O187S12, M+nH, 1212.10991
2000
1500
1000
500
0
1211
1212
1213
1214
m/z

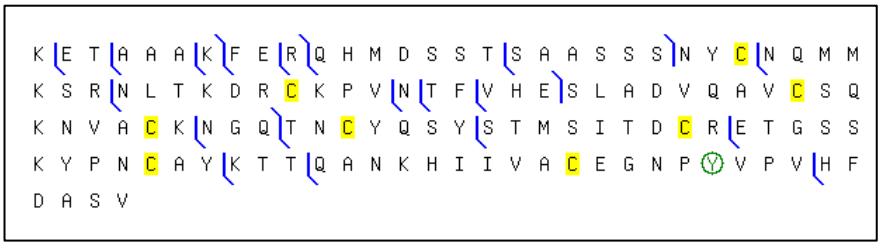

z_{121}^{10+}
Intens.
x10^6
4
2
0
00003.d:C560H882N167O187S12, M+nH, 1333.22017
2000
1500
1000
500
0
1333
1334
1335
m/z


z_{123}^{11+}
Intens.
x10^7
1.0
0.5
0.0
00003.d:C569H896N169O192S12, M+nH, 1233.02720
2000
1500
1000
500
0
1232
1233
1234
1235 m/z

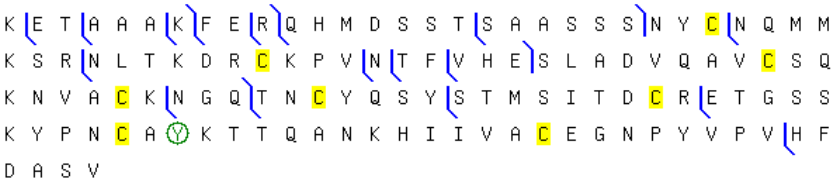
ECD: Assessment of potentially reactive sites

The most complete sequence coverage was observed when the modification was modelled on Y115 using Prosite PTM. Sequence coverage was lost when the modification was modelled on other Tyr residues. Sequence coverage was also lost when the modification was modelled on His residues, which are known to exhibit cross-reactivity with diazonium salts. The single exception to this is H105, which could not be distinguished from Y115 under ECD fragmentation conditions, however the residues were separable by CID fragmentation. When the unmodified protein was modelled, no peptides containing the Y115 residue could be identified.

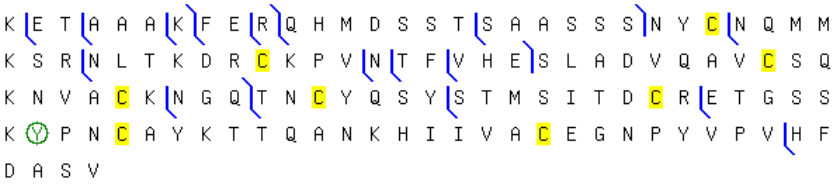
Y115



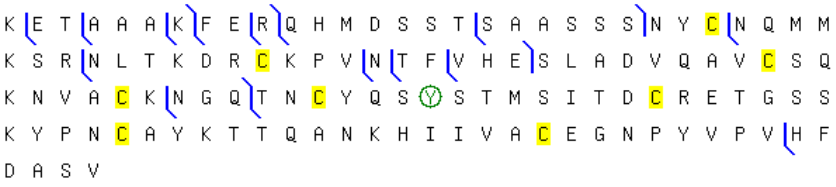
Y097



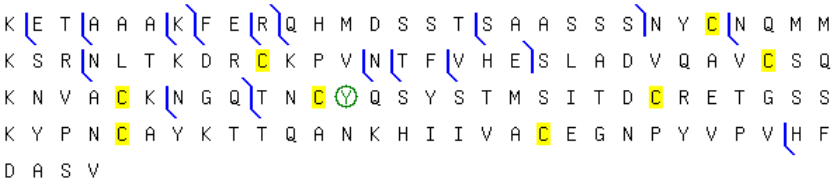
Y092



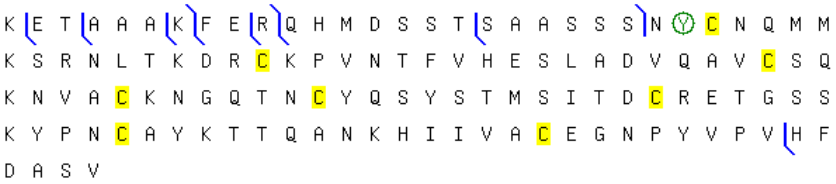
Y076



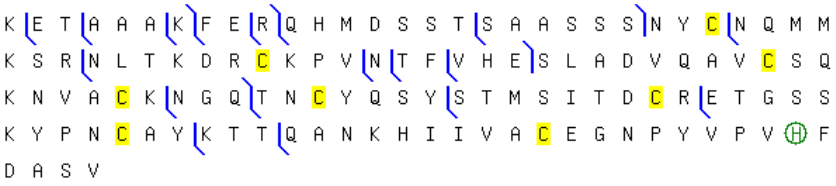
Y073



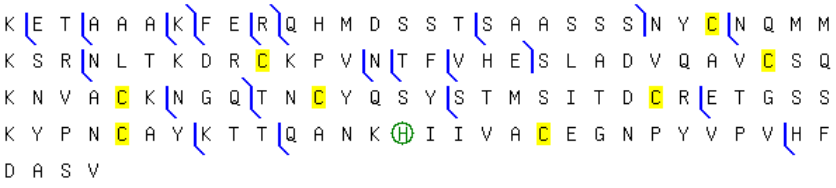
Y025



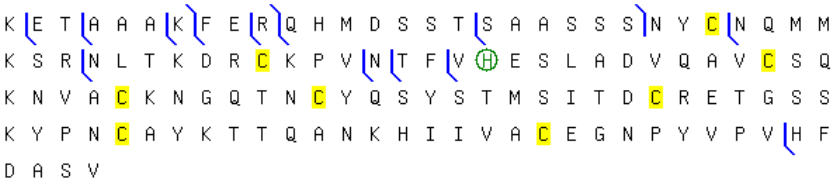
H119



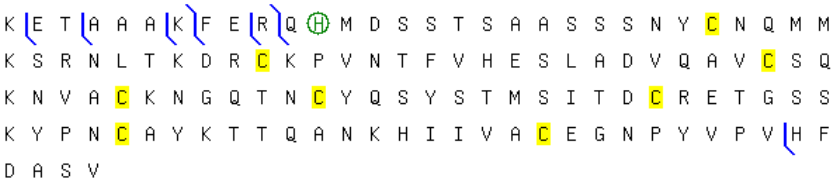
H105



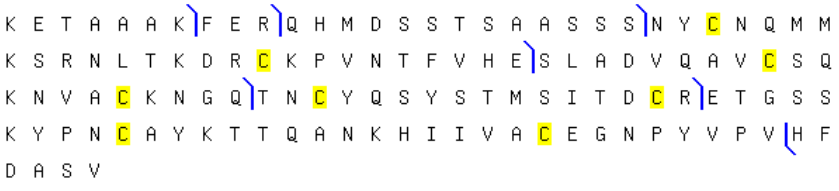
H048



H012



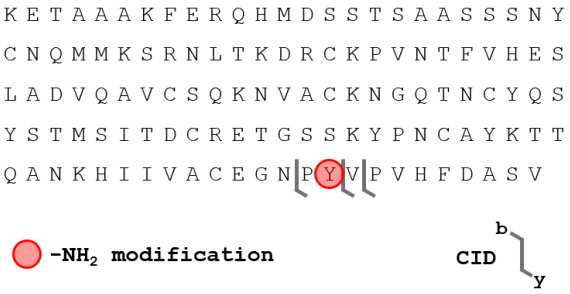
Unmodified



CID fragmentation

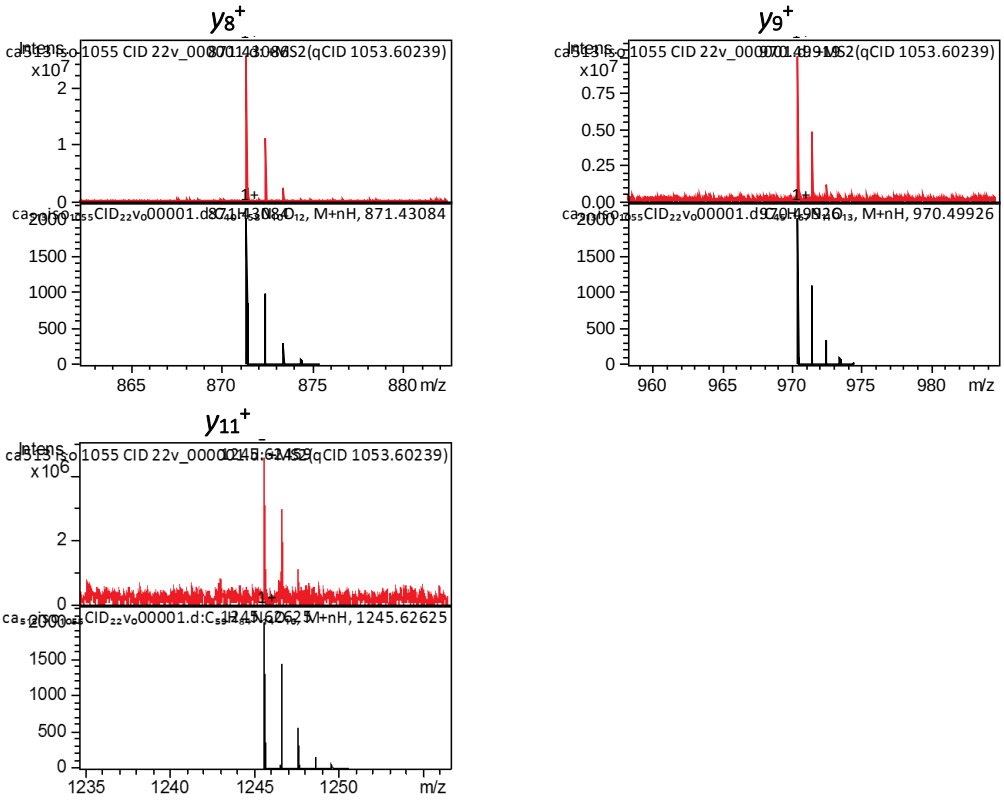
The 13+ charge state of the singly modified protein was isolated with ion accumulation time of up to 3 sec. The collision voltage was 22 V. The resulting peptide fragments were calibrated by single point calibration to $[C_{40}H_{59}N_{10}O_{12}]^+$: 871.43084 *m/z*. Using DataAnalysis software (version 4.1, Bruker Daltronics), a list of monoisotopic masses was generated and this was searched against the theoretical masses of all expected peptides using ProSight PTM software (Kelleher Lab, Northwestern University)[168, 169] with a mass error tolerance of 5 ppm. The identified peptide peaks were further scrutinised by comparison with the theoretical isotope distribution using DataAnalysis.

CID: Identified peptide ions



Peptide ions identified by top-down fragmentation of singly modified RNase A by CID.

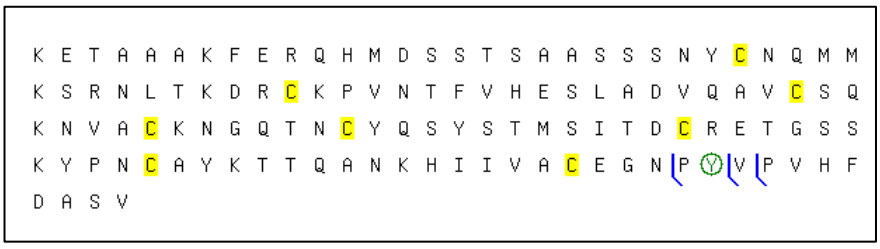
Peptide Ion	*m/z*	*z*	Measured Molecular Mass (Da)	Theoretical Molecular Mass (Da)	Mass Error (ppm)
y8	871.4309	1+	870.4239	870.4255	-1.86
y9	970.4992	1+	969.4922	969.4939	-1.75
y11	1245.6246	1+	1244.6180	1244.6200	-1.58



CID: Assessment of potentially reactive sites

Under CID fragmentation conditions there was poor sequence coverage of the protein, presumably due to insufficient reduction of the four disulphides which are native to RNase A prior to infusion. However, the identified peptides fell within the key C-terminal region (as described by ECD fragmentation), and allowed separation of H105 from Y115. The y_{11} ion contains a –NH_2 modification and includes the Y115 residue. The y_9 and y_8 ions are unmodified and do not include Y115. Some unmodified y_{11} was also identified, and this was thought to arise from the CID conditions which are harsher than those used in ECD.

Y115



Y097

Y092

Y076

Y073

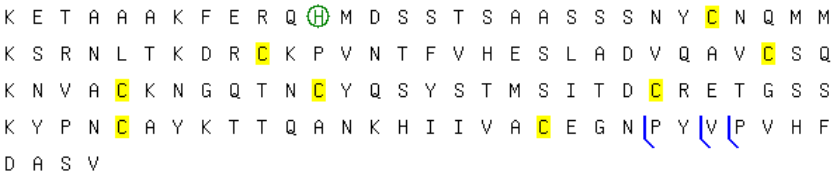
Y025

H119

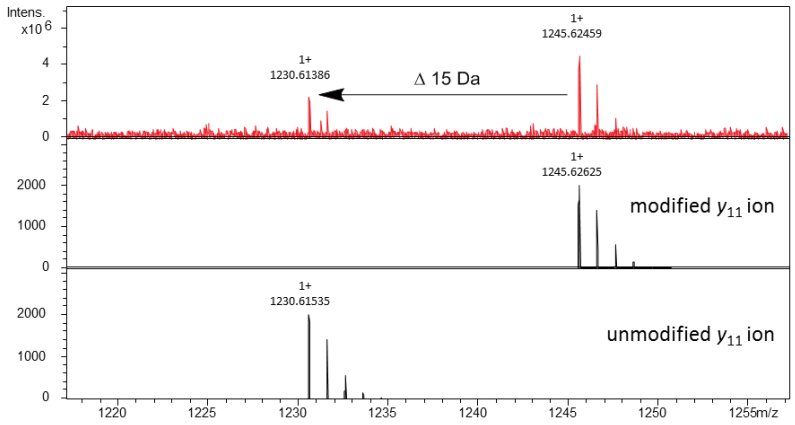
H105

H048

H012



Unmodified



The unmodified y_{11} ion (1230.61535 m/z) was identified by CID, but not by ECD. The modified y_{11} ion (1245.62625 m/z) was also observed. It is assumed that the NH_2 modification was removed under the thermal dissociation conditions in CID fragmentation, but was preserved under milder ECD fragmentation.

5.8.4 Oxidative coupling of catch-and-release modified proteins

To a solution of catch-and-release modified protein (62.5 µM) in phosphate buffer (240 µL, 10 mM, pH 6.5) was added fluoresceinamine (30 µL, 10 mM in DMSO) and $K_3Fe(CN)_6$ (30 µL, 800 mM in buffer). The mixture was briefly mixed by vortex and agitated on a rotator for 18 h at rt. The sample volume was reduced to ~50 µL on a MWCO spin cartridge and desalted by washing with phosphate buffer (5 x 400 µL, 100 mM, pH 7.2).

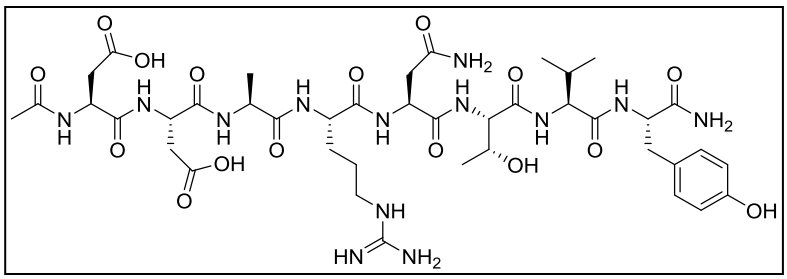
5.8.5 FITC-labelling of RNase A

To a solution of RNase A (290 mM) in bicarbonate buffer (1 mL, 50 mM, pH 9.3) was added fluorescein isothiocyanate (100 µL, 44 mM in DMF). The reaction mixture was briefly mixed by vortex and agitated on a rotator for 30 min at rt. The sample volume was reduced to ~50 µL on a MWCO spin cartridge and desalted by washing with phosphate buffer (5 x 400 µL, 100 mM, pH 7.2).

Appendix 1: Peptides

Peptides synthesised and characterised by project student Christina Burr.

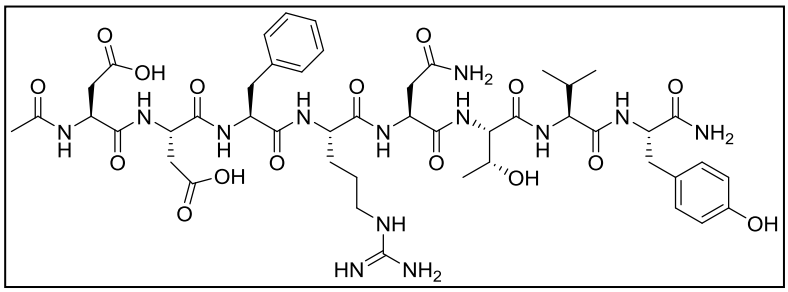
Ac-Asp-Asp-Ala-Arg-Asn-Thr-Val-Tyr-NH_2



The nanobody-derived peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3. The lyophilised peptide was a colourless amorphous solid (96% purity by RP-HPLC at 210 nm, 87% yield).

R_t (HPLC Method 1) = 7.77 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 9.12 (1H, s, Tyr(ArO*H*)), 8.24 (1H, d, *J* = 7.4 Hz, Asp(N*H*)), 8.19 (1H, d, *J* = 7.5 Hz, Asp(N*H*)), 8.14 (1H, d, *J* = 7.5 Hz, Asn(N*H*)), 7.99 (1H, d, *J* = 8.0 Hz, N*H*), 7.75 (1H, d, *J* = 8.2 Hz, N*H*), 7.72 (1H, d, *J* = 8.2 Hz, N*H*), 7.71 (1H, d, *J* = 8.2 Hz, N*H*), 7.68 (1H, d, *J* = 7.0 Hz, N*H*), 7.46 (1H, s, N*H*), 7.34 (1H, s, N*H*), 7.15 (1H, s, N*H*), 7.02 (1H, s, N*H*), 7.00 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 6.63 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 4.98 (1H, s, Thr(O*H*)), 4.63 (1H, dt, *J* = 6.8, 6.8 Hz, Asn(α-C*H*), 4.55-4.46 (2H, m, Asp(α-C*H*), Asp(α-C*H*)), 4.36-4.22 (4H, m, Ala(α-C*H*), Arg(α-C*H*), Thr(α-C*H*), Tyr(α-C*H*)), 4.11-4.00 (2H, m, Thr(β-C*H*), Val(α-C*H*)), 3.13-3.03 (2H, m, Arg(δ-CH_2)), 2.90 (1H, dd, *J* = 13.9, 5.3, Tyr(β-CH_AH$_B$)), 2.76-2.51 (4H, m, Asp(β-CH_AH$_B$), Asp(β-CH_AH$_B$), Asn(β-CH_AH$_B$), Tyr(β-CH$_A$$H_B$)), 2.51-2.43 (3H, m, Asp(β-CH$_A$$H_B$), Asp(β-CH$_A$$H_B$), Asn(β-CH$_A$$H_B$)), 2.03-1.90 (1H, m, Val(β-C*H*), 1.85 (3H, s, COCH_3), 1.74-1.67 (1H, m, Arg(β-CH_AH$_B$)), 1.58-1.43 (3H, m, Arg(β-CH$_A$$H_B$, γ-C$H_2$)), 1.23 (3H, d, *J* = 7.0 Hz, Ala(β-CH_3), 1.04 (3H, d, *J* = 6.3 Hz, Thr(CH_3)), 0.74 (3H, d, *J* = 6.9 Hz, Val(CH_3), 0.74 (3H, d, *J* = 6.9 Hz, Val(CH_3); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.43 (*C*O), 172.51 (*C*O), 172.44 (*C*O), 172.23 (*C*O), 172.22 (*C*O), 171.55 (*C*O), 171.53 (*C*O), 171.43 (*C*O), 170.96 (*C*O), 170.55 (*C*O), 170.50 (*C*O), 170.25 (*C*O), 157.03 (Ar*C*), 156.21 (Ar*C*), 130.41 (2Ar*C*H), 128.32 (*C*N), 115.35 (2Ar*C*H), 66.88 (*C*H), 58.90 (*C*H), 58.52 (*C*H), 54.74 (*C*H), 52.34 (*C*H), 50.23 (*C*H), 50.20 (*C*H), 49.91 (*C*H), 48.86 (*C*H), 40.92 (*C*H_2), 37.49 (*C*H_2), 37.12 (*C*H_2), 36.38 (*C*H_2), 36.17 (*C*H_2), 30.52 (*C*H), 29.60 (*C*H_2), 25.27 (*C*H_3), 22.98 (*C*H_3), 19.89 (*C*H_3), 19.55 (*C*H_3), 18.34 (2*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 995 ([M+H]$^+$, 15%), 509 ([M+H+Na]$^{2+}$, 100), 498 ([M+2H]$^{2+}$, 35); **HRMS** (ESI+, MeCN/H_2O) [M+2H]$^{2+}$ found 497.7327, [$C_{41}H_{65}N_{13}O_{16}$]$^{2+}$ requires 497.7331.

Ac-Asp-Asp-Phe-Arg-Asn-Thr-Val-Tyr-NH_2

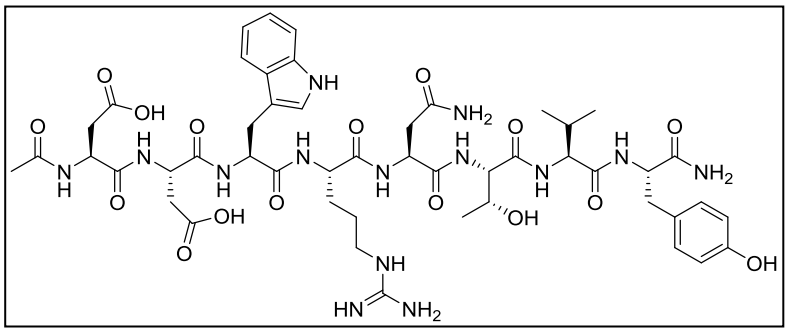


The Phe-substituted peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3.

The lyophilised peptide was a colourless amorphous solid (98% purity by RP-HPLC at 210 nm, 79% yield).

R_t (HPLC Method 1) = 8.28 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 9.47 (1H, s, N*H*), 9.21 (1H, s, Tyr(ArO*H*)), 8.64 (1H, d, *J* = 6.9 Hz, N*H*), 8.56 (1H, d, *J* = 6.3 Hz, N*H*), 8.10 (1H, s, N*H*), 8.09 (1H, s, N*H*), 7.94 (1H, s, N*H*), 7.73-7.53 (4H, m, 4 x N*H*), 7.27-7.20 (4H, m, Phe(4Ar*H*)), 7.19-7.13 (1H, m, Phe(Ar*H*)), 7.08-7.02 (2H, m, 2 x N*H*), 7.01 (2H, d, *J* = 8.4 Hz, Tyr(2Ar*H*)), 6.94 (1H, s, 2 x N*H*), 6.64 (2H, d, *J* = 8.4 Hz, Tyr(2Ar*H*)), 5.12 (1H, s, Thr(O*H*)), 4.59 (1H, q, *J* = 7.4 Hz, α-C*H*), 4.50 (1H, q, *J* = 7.0 Hz, Asp(α-C*H*)), 4.44 (1H, td, *J* = 8.8, 3.6 Hz, α-C*H*), 4.34 (1H, ddd, *J* = 10.4, 6.8, 3.5 Hz, α-C*H*), 4.29 (1H, td, *J* = 8.9, 4.9 Hz, α-C*H*), 4.21 (1H, dd, *J* = 7.8, 4.1 Hz, Thr(α-C*H*)), 4.10-4.02 (2H, m, Arg(α-C*H*), Thr(β-C*H*)), 4.02-3.97 (1H, m, Val(α-C*H*)), 3.14 (1H, dd, *J* = 14.2, 3.5 Hz, β-CH_AH_B), 3.11-3.06 (1H, m, Arg(δ-CH_AH_B)), 3.03-2.95 (2H, m, Arg(δ-CH_A*H_B*), β-C$H_C$$H_D$), 2.93 (1H, dd, *J* = 14.2, 4.9 Hz, β-CH_A*H_B*), 2.71-2.61 (2H, m, β-CH_C*H_D*, β-C$H_E$$H_F$), 2.59-2.52 (1H, m, β-C$H_E$C*$H_F$*), 2.50-2.43 (2H, m, Asp(β-CH_ACH$_B$), β-CH_GCH$_H$), 2.38 (1H, dd, *J* = 16.0, 7.0 Hz, Asp(β-CH_AC*H_B*)), 2.26 (1H, dd, *J* = 15.5, 3.5 Hz, β-CH_GC*H_H*), 1.99-1.89 (1H, m, Val(β-C*H*)), 1.82 (3H, s, COCH_3), 1.73-1.67 (2H, m, Arg(β-CH_2)), 1.61-1.46 (2H, m, Arg(γ-CH_2)), 1.06 (3H, d, *J* = 6.3 Hz, Thr(CH_3), 0.71 (3H, d, *J* = 6.8 Hz, Val(CH_3)), 0.70 (3H, d, *J* = 6.8 Hz, Val(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.52 (*C*O), 173.07 (*C*O), 173.02 (*C*O), 172.92 (*C*O), 172.45 (*C*O), 172.24 (*C*O), 172.05 (*C*O), 171.81 (*C*O), 170.90 (*C*O), 170.84 (*C*O), 170.76 (*C*O), 169.74 (*C*O), 157.90 (Ar*C*), 156.23 (Ar*C*), 138.65 (Ar*C*), 130.36 (2Ar*C*H), 129.42 (2Ar*C*H), 128.56 (2Ar*C*H), 128.40 (*C*N), 126.59 (Ar*C*H), 115.39 (2Ar*C*H), 67.14 (*C*H), 59.24 (*C*H), 59.15 (*C*H), 55.81 (*C*H), 54.95 (*C*H), 54.38 (*C*H), 50.67 (*C*H), 50.56 (*C*H), 50.02 (*C*H), 41.10 (*C*H_2), 38.07 (*C*H_2), 37.25 (*C*H_2), 36.91 (*C*H_2), 36.08 (*C*H_2), 30.25 (*C*H), 28.29 (*C*H_2), 24.22 (*C*H_2), 22.94 (*C*H_3), 19.91 (*C*H_3), 19.45 (*C*H_3), 18.30 (*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 1071 ($[M+H]^+$, 20%), 555 ($[M+H+K]^{2+}$, 50), 547 ($[M+H+Na]^{2+}$, 100), 536 ($[M+2H]^{2+}$, 40); **HRMS** (ESI+, MeCN/H_2O) $[M+2H]^{2+}$ found 535.7474, $[C_{47}H_{69}N_{13}O_{16}]^{2+}$ requires 535.7487.

Ac-Asp-Asp-Trp-Arg-Asn-Thr-Val-Tyr-NH_2,

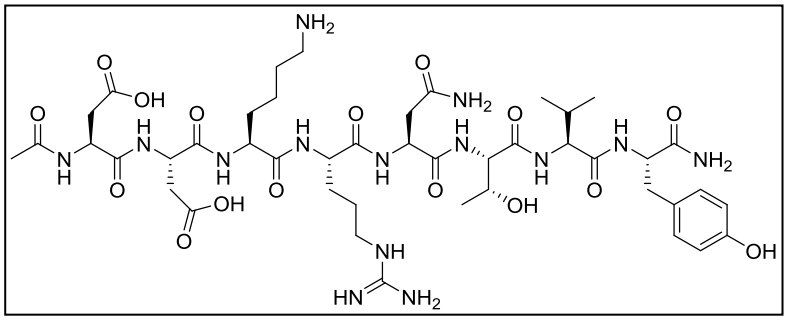


The Trp-substituted peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3. The product was purified by preparative RP-HPLC (System 4) using an isocratic solvent system (H_2O + 0.1%: MeCN + 0.1% TFA, 94:6). The lyophilised peptide was a colourless amorphous solid (96% purity by RP-HPLC at 210 nm, 31% yield).

R_t (HPLC Method 1) = 8.39 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 12.34 (2H, br s, 2 x Asp(CO_2*H*)), 10.72 (1H, d, *J* = 2.5 Hz, Trp(Indole-N*H*)), 9.14 (1H, s, Tyr(ArO*H*)), 8.22 (2H, d, *J* = 7.5 Hz, 2 x Asp(N*H*)), 8.20-8.13 (1H, m, N*H*), 8.05 (2H, br s, 2 x N*H*), 7.76 (1H, d, *J* = 8.2 Hz, N*H*), 7.74-7.68 (2H, m, Thr(N*H*), Val(N*H*)), 7.55 (1H, d, *J* = 7.9 Hz, N*H*), 7.49 (1H, s, N*H*), 7.32 (1H, d, *J* = 8.0 Hz, N*H*), 7.16 (1H, br s, Trp(Ar*H*)), 7.09-6.93 (6H, m, Trp(4Ar*H*), Tyr(2Ar*H*)), 6.63 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*), 5.00 (1H, d, *J* = 5.3 Hz, Thr(O*H*), 4.67 (1H, q, *J* = 7.0 Hz, α-C*H*), 4.59-4.40 (3H, m, α-C*H*, 2 x Asp(α-C*H*)), 4.32 (1H, td, *J* = 8.7, 5.3 Hz, α-C*H*), 4.26 (1H, dd, *J* = 8.1, 3.8 Hz, Thr(α-C*H*)), 4.09-4.03 (2H, m, Thr(β-C*H*), Val(α-C*H*)), 3.15 (1H, dd, *J* = 14.9, 4.1 Hz, β-C$H_A$$H_B$), 3.10-2.95 (3H, m, β-C$H_A$$H_B$, Arg(δ-C$H_2$)), 2.90 (1H, dd, *J* = 13.9, 5.2 Hz, β-C$H_C$$H_D$), 2.72-2.59 (4H, m, β-C$H_C$$H_D$, β-C$H_E$$H_F$, 2 x Asp(β-C$H_A$$H_B$)), 2.57-2.39 (3H, m, β-C$H_E$$H_F$, 2 x Asp(β-C$H_A$$H_B$)), 2.03-1.90 (1H, m, Val(β-C*H*)), 1.83 (3H, s, COCH_3), 1.74-1.62 (1H, m, Arg(β-C$H_A$$H_B$)), 1.60-1.48 (1H, m, Arg(β-C$H_A$$H_B$)), 1.48-1.32 (2H, m, Arg(γ-CH_2)), 1.05 (3H, d, *J* = 6.3 Hz, Thr(CH_3)), 0.74 (3H, d, *J* = 6.8 Hz, Val(CH_3)), 0.73 (3H, d, *J* = 6.8 Hz, Val(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.44 (*C*O), 172.31 (*C*O), 172.24 (*C*O), 171.64 (*C*O), 171.59 (*C*O), 170.96 (*C*O), 170.53 (*C*O), 170.27 (*C*O), 158.32 (Ar*C*), 158.08 (Ar*C*), 157.12 (Ar*C*), 156.21 (*C*O), 145.12 (Ar*C*), 136.44 (*C*O), 130.41 (2Ar*C*H), 128.33 (*C*O), 127.79 (*C*O), 124.21 (Ar*C*H), 121.26 (Ar*C*H), 118.81 (Ar*C*H), 118.69 (Ar*C*H), 116.65 (*C*N), 115.35 (2Ar*C*H), 111.70 (Ar*C*H), 110.22 (Ar*C*), 66.95 (*C*H), 58.94 (*C*H), 58.58 (*C*H), 54.76 (*C*H), 54.24 (*C*H), 50.24 (*C*H), 50.23 (*C*H), 50.16 (*C*H), 40.97 (*C*H_2), 37.46 (*C*H_2), 37.11 (2*C*H_2), 36.39 (*C*H_2), 30.50 (*C*H), 29.53 (*C*H_2), 27.59 (*C*H_2), 25.00 (*C*H_2), 22.96 (*C*H_3), 19.90 (*C*H_3), 19.54 (*C*H_3), 18.35 (*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 1110 ([M+H]$^+$, 25%), 574 ([M+H+K]$^{2+}$, 80), 566 ([M+H+Na]$^{2+}$, 100), 555 ([M+2H]$^{2+}$, 70);

HRMS (ESI, MeCN/H_2O) $[M+H+K]^{2+}$ found 574.2310, $[C_{49}H_{69}N_{14}O_{16}K]^{2+}$ requires 574.2321.

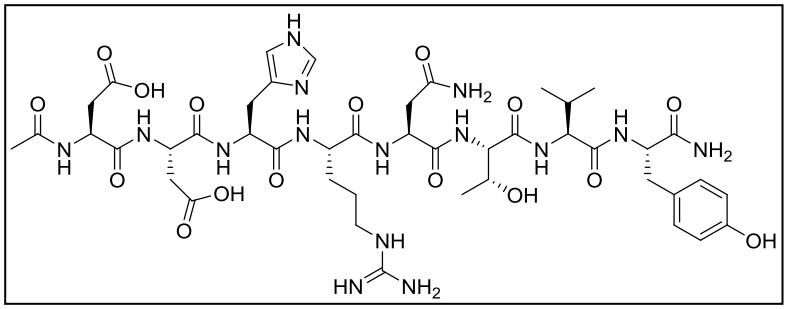
Ac-Asp-Asp-Lys-Arg-Asn-Thr-Val-Tyr-NH_2



The Lys-substituted peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3. The lyophilised peptide was a colourless amorphous solid (96% purity by RP-HPLC at 210 nm).

R_t (HPLC Method 1) = 7.77 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 11.41 (2H, br s, 2 x Asp(CO_2*H*)), 9.15 (1H, s, Tyr(ArO*H*)), 8.46 (1H, br s, N*H*), 8.25 (1H, d, *J* = 7.6 Hz, Asp(N*H*)), 8.04 (2H, br s, 2 x N*H*), 7.75-7.67 (3H, Lys(N*H*), Arg(N*H*), N*H*), 7.69 (1H, d, *J* = 8.1 Hz, Val(N*H*)), 7.62 (1H, d, *J* = 8.1 Hz, Thr(N*H*)), 7.48 (1H, s, (N*H*), 7.18-6.90 (6H, m, Arg(2N*H*), Tyr(2Ar*H*), 2N*H*), 6.64 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 4.99 (1H, s, Thr(O*H*)), 4.62 (1H, q, *J* = 7.2 Hz, α-C*H*), 4.52 (1H, dd, *J* = 13.7, 7.7 Hz, Asn(α-C*H*)), 4.50-4.43 (1H, m, α-CH), 4.31 (1H, td, *J* = 8.7, 5.1 Hz, α-CH), 4.24 (1H, dd, *J* = 8.1, 3.9 Hz, Thr(α-CH)), 4.13 (2H, br s, Lys(α-C*H*), Arg(α-C*H*)), 4.07-4.01 (2H, m, Val(α-C*H*), Thr(β-C*H*)), 3.14-3.08 (1H, m, β-C$H_A$$H_B$), 3.06-2.99 (1H, m, β-C$H_A$$H_B$), 2.91 (1H, dd, *J* = 14.0, 5.1, β-C$H_C$$H_D$), 2.81-2.76 (1H, m, C$H_C$$H_D$), 2.68-2.61 (4H, m, Asp(β-C$H_A$$H_B$), β-C$H_E$$H_F$, β-C$H_G$$H_H$, β-C$H_G$$H_H$), 2.49-2.41 (2H, m, Asp(β-C$H_A$$H_B$) β-C$H_E$$H_F$), 1.99-1.90 (1H, m, Val(β-C*H*)), 1.85 (3H, s, COCH_3), 1.80-1.64 (3H, m, Arg/Lys CH_2), 1.65-1.44 (5H, m, Arg/Lys CH_2), 1.42-1.29 (2H, m, Lys(CH_2)), 1.05 (3H, d, *J* = 6.4 Hz, Thr(CH_3)), 0.73 (3H, d, *J* = 7.1 Hz, Val(CH_3)), 0.72 (3H, d, *J* = 7.1 Hz, Val(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.46 (*C*O), 173.36 (*C*O), 172.52 (*C*O), 172.50 (*C*O), 172.10 (*C*O), 172.04 (*C*O), 171.59 (*C*O), 170.93 (*C*O), 170.79 (*C*O), 170.54 (*C*O), 170.25 (*C*O), 158.34 (Ar*C*), 157.55 (Ar*C*), 156.21 (*C*O), 130.39 (2Ar*C*H), 128.34 (*C*N), 115.36 (2Ar*C*H), 67.02 (*C*H), 60.22 (*C*H), 58.98 (*C*H), 58.70 (*C*H), 54.80 (*C*H), 53.58 (*C*H), 50.54 (*C*H), 50.39 (*C*H), 50.15 (*C*H), 41.05 (*C*H_2), 38.95 (*C*H_2), 37.39 (*C*H_2), 37.05 (*C*H_2), 36.76 (*C*H_2), 31.42 (*C*H_2), 30.54 (*C*H_2), 30.43 (*C*H), 28.77 (*C*H_2), 26.87 (*C*H_2), 24.63 (*C*H_2), 22.96 (*C*H_3), 22.54 (*C*H_2), 19.87 (*C*H_3), 19.52 (*C*H_3), 18.29 (*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 1052 ($[M+H]^+$, 5%), 526 ($[M+2H]^{2+}$, 100); **HRMS** (ESI+, MeCN/H_2O) $[M+H]^+$ found 1051.5184, $C_{44}H_{71}N_{14}O_{16}$ requires 1051.5167.

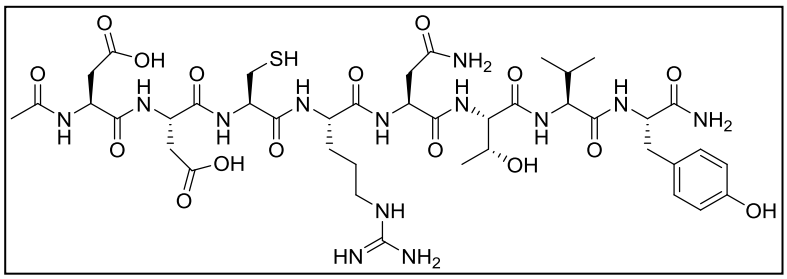
Ac-Asp-Asp-His-Arg-Asn-Thr-Val-Tyr-NH_2



The Phe-substituted peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3. The lyophilised peptide was a colourless amorphous solid (96% purity by RP-HPLC at 210 nm).

R_t (HPLC Method 1) = 7.80 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 8.51 (1H, s, N*H*), 8.26-8.20 (1H, m, Arg(N*H*)), 8.23 (1H, d, *J* = 7.5 Hz, Asp(N*H*)), 8.19 (1H, d, *J* = 7.9 Hz, Asp(N*H*)), 7.94 (1H, s, N*H*), 7.76 (1H, d, *J* = 8.2 Hz, N*H*), 7.72 (1H, d, *J* = 8.2 Hz, Val(N*H*)), 7.69 (1H, d, *J* = 7.9 Hz, Thr(N*H*)), 7.51 (1H, s, His(ArN*H*)), 7.16 (1H, s, N*H*), 7.10 (1H, s, N*H*), 7.04 (1H, s, N*H*), 7.02-6.97 (4H, m, His(2Ar*H*), Tyr(2Ar*H*)), 6.63 (2H, d, *J* = 8.5 Hz, Tyr(2Ar*H*)), 4.66 (1H, q, *J* = 7.1 Hz, α-C*H*), 4.59-4.40 (3H, m, 2 x Asp(α-C*H*), α-C*H*), 4.32 (1H, td, *J* = 8.7, 5.3 Hz, α-C*H*), 4.25 (1H, dd, *J* = 8.0, 4.0 Hz, Thr(α-C*H*)), 4.24-4.20 (1H, m, α-C*H*), 4.10-4.00 (2H, m, Thr(β-C*H*), Val(α-C*H*)), 3.10-3.04 (3H, m, Arg(δ-CH_2), β-C$H_A$$H_B$), 2.97 (1H, dd, *J* = 15.0, 8.1 Hz, CH_AH_B), 2.90 (1H, dd, *J* = 14.0, 5.2 Hz, β-C$H_C$$H_D$), 2.71-2.60 (4H, m, 2 x Asp(C$H_A$$H_B$), β-$CH_C$$H_D$, β-C$H_E$$H_F$), 2.51-2.43 (3H, m, 2 x Asp(β-CH_AH_B), β-CH_EH_F), 1.99-1.90 (1H, m, Val(β-C*H*)), 1.85 (3H, s, CH_3CO), 1.75-1.69 (1H, m, Arg(β-C$H_A$$H_B$)), 1.61-1.55 (1H, m, Arg(β-CH_AH_B)), 1.53-1.46 (2H, m, Arg(γ-CH_2)), 1.03 (3H, d, *J* = 6.3 Hz, Thr(CH_3)), 0.73 (3H, d, *J* = 6.8 Hz, Val(CH_3)), 0.72 (3H, d, *J* = 6.8 Hz, Val(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.44 (*C*O), 172.82 (*C*O), 172.45 (*C*O), 172.24 (*C*O), 171.87 (*C*O), 171.58 (*C*O), 171.45 (*C*O), 171.04 (*C*O), 170.95 (*C*O), 170.74 (*C*O), 170.49 (*C*O), 170.22 (*C*O), 158.40 (Ar*C*), 157.21 (Ar*C*), 156.21 (Ar*C*), 134.68 (Ar*C*H), 130.40 (2Ar*C*H), 128.32 (*C*N), 117.18 (Ar*C*H), 115.35 (2Ar*C*H), 66.89 (*C*H), 60.22 (*C*H), 58.88 (*C*H), 58.55 (*C*H), 54.74 (*C*H), 53.06 (*C*H), 52.88 (*C*H), 50.33 (*C*H), 50.20 (*C*H), 40.96 (*C*H), 37.40 (*C*H_2), 37.11 (*C*H_2), 36.68 (*C*H_2), 30.51 (*C*H_2), 29.24 (*C*H_2), 25.00 (*C*H_2), 22.99 (*C*H_2), 22.53 (*C*H_2), 19.87 (*C*H_3), 19.54 (*C*H_3), 18.33 (*C*H_3), 14.56 (*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 531 ($[M+2H]^{2+}$, 100%); **HRMS** (ESI+, MeCN/H_2O) $[M+2H]^{2+}$ found 530.7457, $[C_{44}H_{67}N_{15}O_{16}]^{2+}$ requires 530.7440.

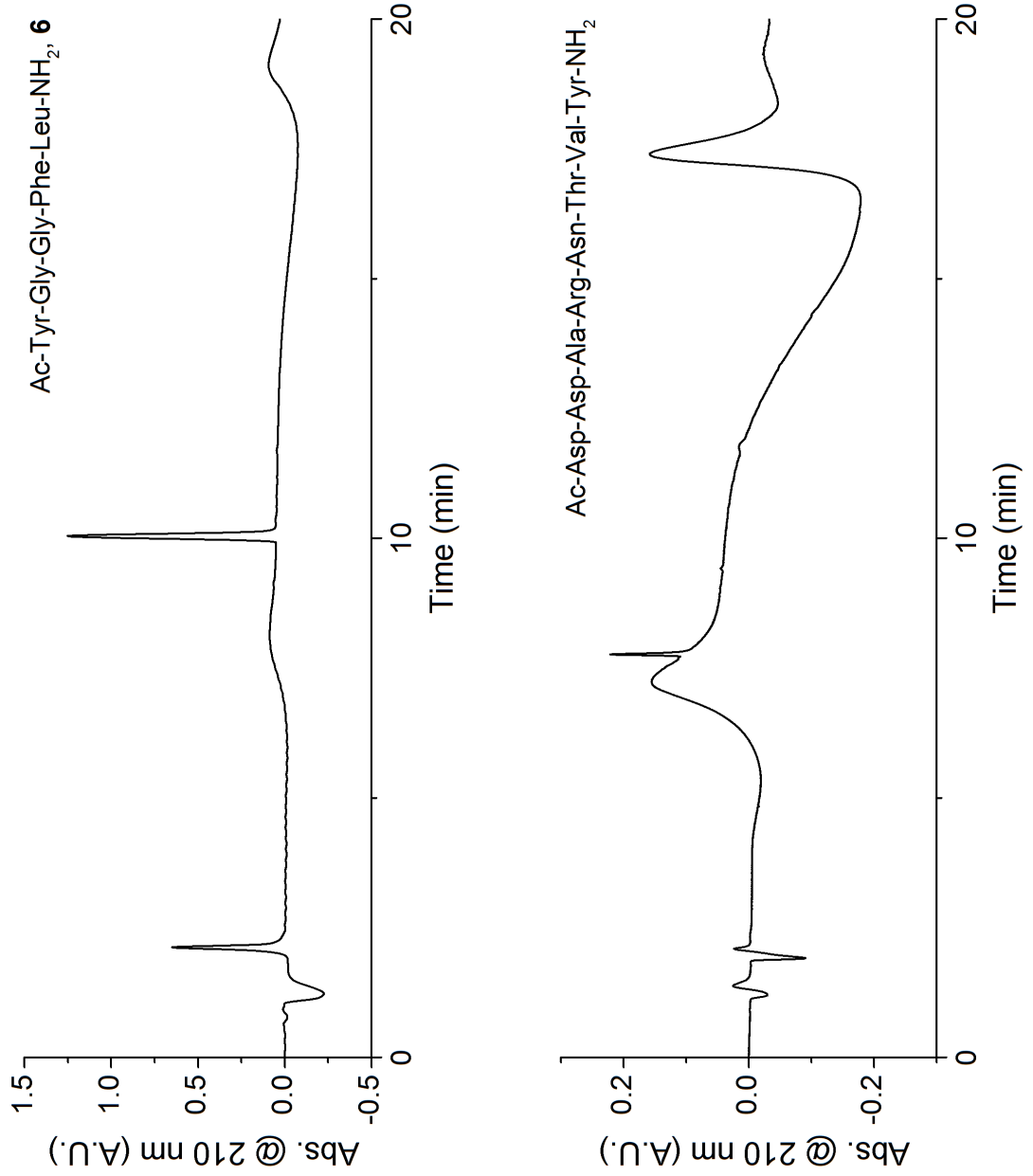
Ac-Asp-Asp-Cys-Arg-Asn-Thr-Val-Tyr-NH_2

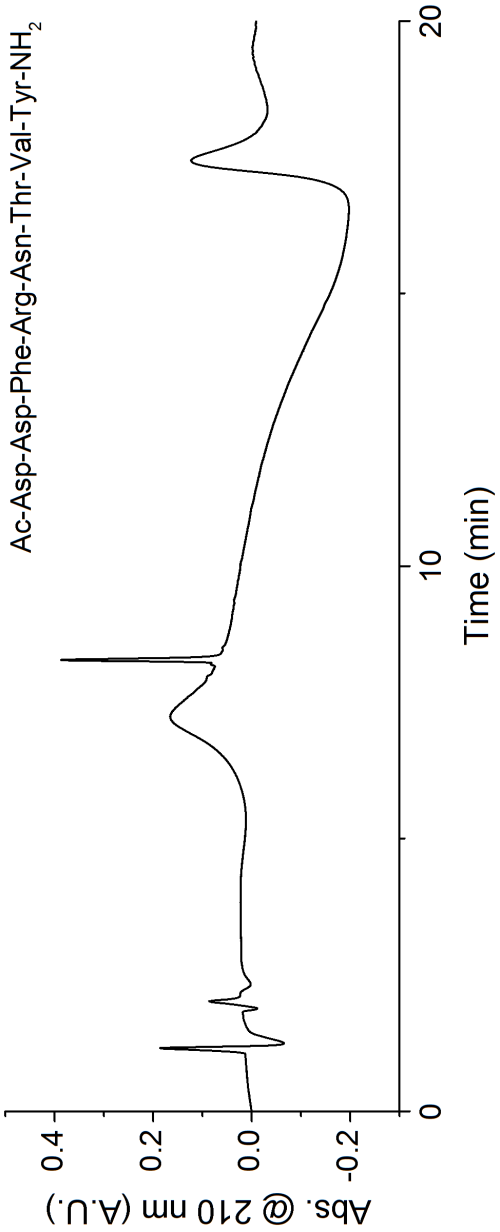


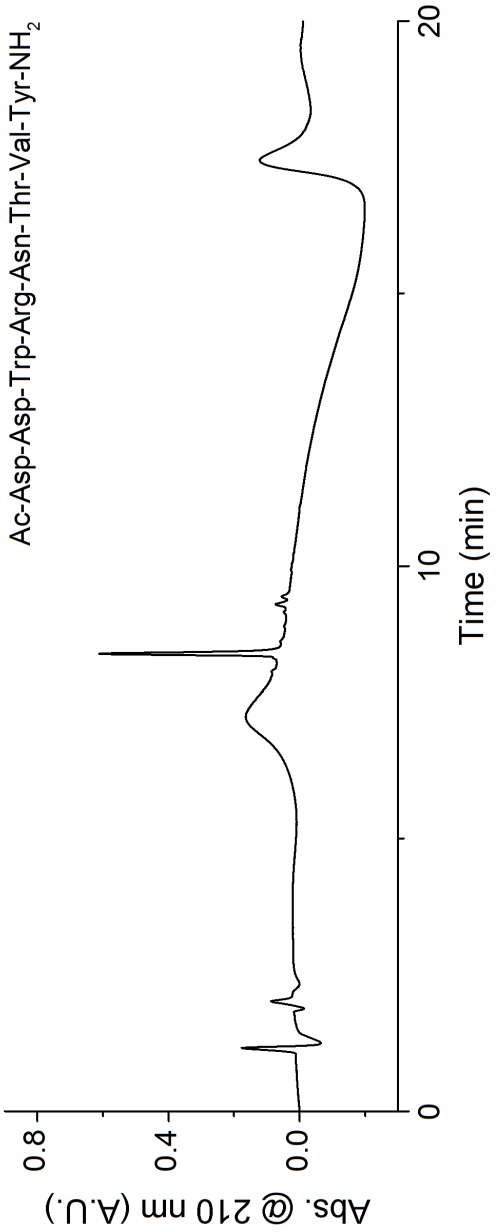
The Phe-substituted peptide was synthesised according to the solid-phase peptide synthesis method described in Section 5.3. The product was purified by preparative RP-HPLC (System 4) using an isocratic solvent system (H_2O + 0.1%: MeCN + 0.1% TFA, 87:13). The lyophilised peptide was a colourless amorphous solid (99% purity by RP-HPLC at 210 nm).

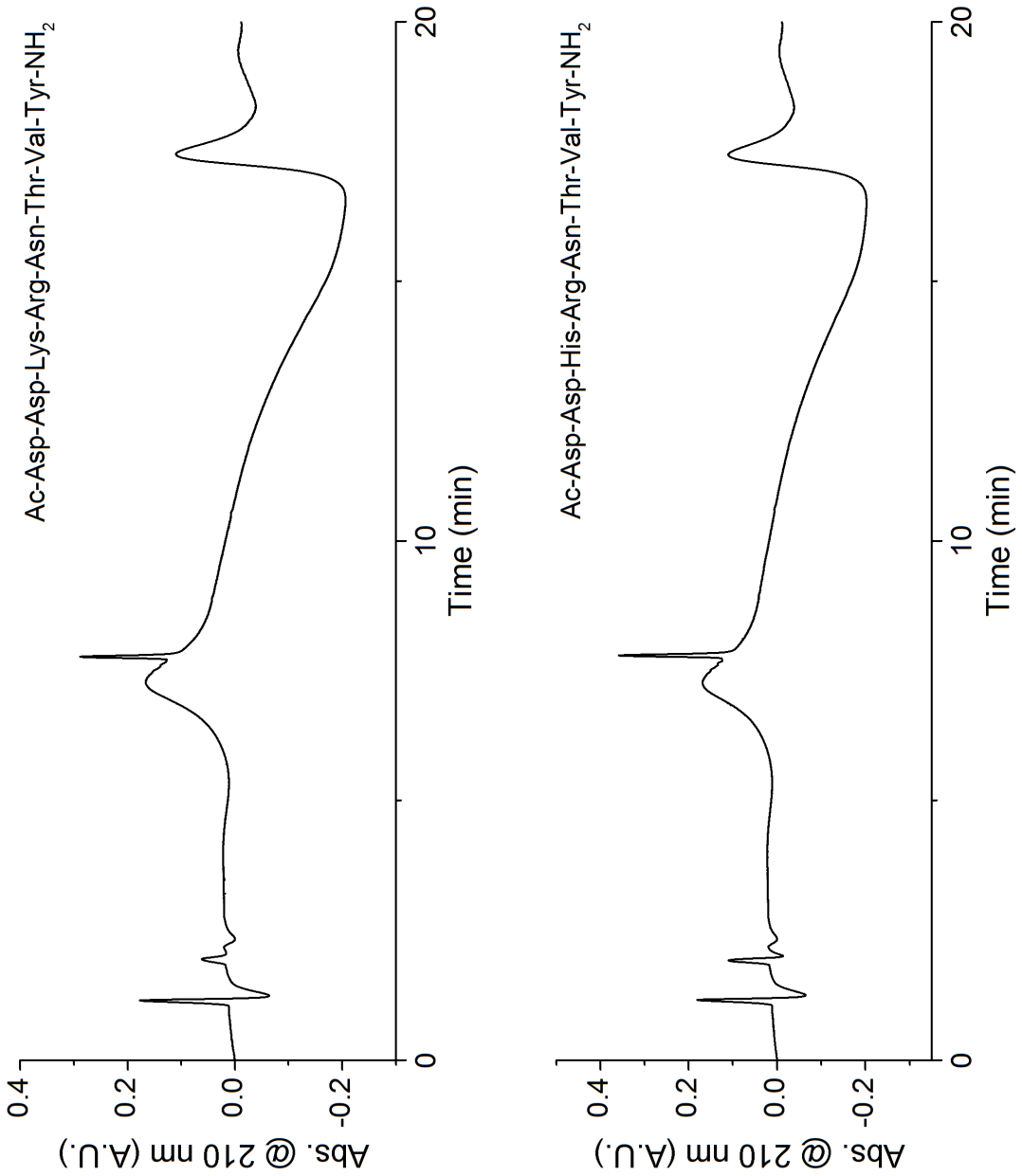
R_t (HPLC Method 1) = 7.84 min; **^{1}H NMR** δ (500 MHz, DMSO-*d_6*) 12.35 (2H, br s, 2 x Asp(CO_2*H*)), 9.14 (1H, s, Tyr(ArO*H*)), 8.26 (2H, d, *J* = 7.2 Hz, 2 x Asp(N*H*)), 8.17 (1H, s, N*H*), 7.77-7.71 (4H, m, Arg(N*H*), Thr(N*H*), Val(N*H*), N*H*), 7.47 (1H, s, N*H*), 7.15 (1H, s, N*H*), 7.03 (1H, s, N*H*), 7.02-6.97 (3H, m, Tyr(2Ar*H*), N*H*), 6.63 (2H, d, *J* = 8.6 Hz, Tyr(2Ar*H*)), 4.99 (1H, d, *J* = 5.2 Hz, Thr(O*H*)), 4.63 (1H, q, *J* = 6.9 Hz, α-C*H*), 4.58-4.47 (2H, m, 2 x Asp(α-C*H*)), 4.38 (1H, q, *J* = 6.9 Hz, α-C*H*), 4.31 (1H, td, *J* = 8.6, 5.3 Hz, Arg(α-C*H*)), 4.24 (1H, dd, *J* = 8.1, 3.8 Hz, Thr(α-C*H*)), 4.08-4.03 (2H, m, Thr(β-C*H*), Val(α-C*H*)), 3.10-3.05 (2H, m, Arg(δ-CH_2)), 2.90 (1H, dd, *J* = 13.9, 5.2 Hz, β-C$H_A$$H_B$), 2.85-2.73 (1H, m, β-C$H_C$$H_D$), 2.70-2.58 (4H, m, 2 x Asp(β-C$H_A$$H_B$), $CH_A$$H_B$, C$H_E$$H_F$), 2.58-2.43 (3H, m, 2 x Asp(β-$CH_A$$H_B$), β-$CH_E$$H_F$), 2.41-2.29 (1H, m, $CH_C$$H_D$), 2.03-1.89 (1H, m, Val(β-C*H*)), 1.85 (1H, s, CH_3CO), 1.77-1.65 (1H, m, (Arg(β-C$H_A$$H_B$)), 1.64-1.42 (3H, m, Arg(β-$CH_A$$H_B$), Arg(γ-C$H_2$)), 1.04 (3H, d, *J* = 6.3 Hz, Thr(CH_3)), 0.74 (3H, d, *J* = 6.9 Hz, Val(CH_3)), 0.73 (3H, d, *J* = 6.9 Hz, Val(CH_3)); **^{13}C NMR** δ (126 MHz, DMSO-*d_6*) 173.45 (*C*O), 173.43 (*C*O), 172.29 (*C*O), 172.26 (*C*O), 172.18 (*C*O), 171.64 (*C*O), 171.54 (*C*O), 171.18 (*C*O), 170.96 (*C*O), 170.52 (*C*O), 170.36 (*C*O), 157.16 (Ar*C*), 156.45 (Ar*C*), 156.21 (*C*O), 130.41 (2Ar*C*H), 128.33 (*C*N), 115.35 (2Ar*C*H), 66.88 (*C*H), 58.93 (*C*H), 58.54 (*C*H), 55.77 (*C*H), 54.75 (*C*H), 52.79 (*C*H), 50.30 (*C*H), 50.28 (*C*H), 50.25 (*C*H), 40.95 (*C*H_2), 37.50 (*C*H_2), 37.11 (*C*H_2), 36.40 (*C*H_2), 30.50 (*C*H), 29.41 (*C*H_2), 26.48 (*C*H_2), 25.24 (2*C*H_2), 22.99 (*C*H_3), 19.89 (*C*H_3), 19.54 (*C*H_3), 18.36 (*C*H_3); ***m/z*** (ESI+, MeCN/H_2O) 1048 ([M+Na]$^+$, 25%), 1026 ([M+H]$^+$, 100), 525 ([M+2H]$^{2+}$, 15); **HRMS** (ESI+, MeCN/H_2O) [M+Na]$^+$ found 1048.4099, $C_{41}H_{63}N_{13}O_{16}SNa$ requires 1048.4129.

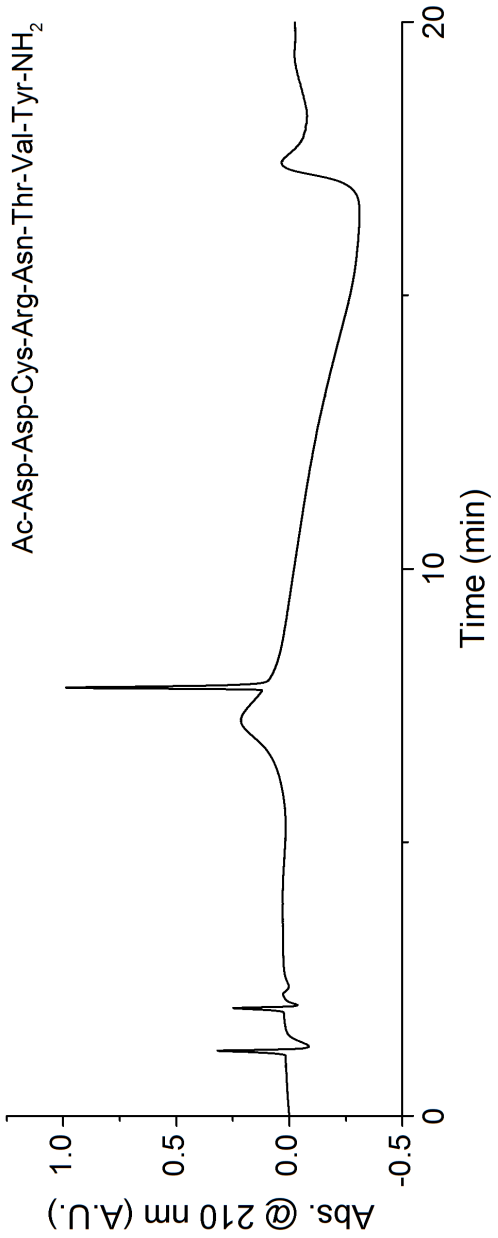
Appendix 2: Peptide HPLC data




Ac-Asp-Asp-Phe-Arg-Asn-Thr-Val-Tyr-NH2
Abs. @ 210 nm (A.U.)
0.4
0.2
0.0
-0.2
0
10
20
Time (min)


Ac-Asp-Asp-Trp-Arg-Asn-Thr-Val-Tyr-NH2
Abs. @ 210 nm (A.U.)
0.8
0.4
0.0
0
10
20
Time (min)


Ac-Asp-Asp-Lys-Arg-Asn-Thr-Val-Tyr-NH2
Abs. @ 210 nm (A.U.)
0.4
0.2
0.0
-0.2
0
10
20
Time (min)
Ac-Asp-Asp-His-Arg-Asn-Thr-Val-Tyr-NH2
Abs. @ 210 nm (A.U.)
0.4
0.2
0.0
-0.2
0
10
20
Time (min)


Ac-Asp-Asp-Cys-Arg-Asn-Thr-Val-Tyr-NH2
Abs. @ 210 nm (A.U.)
1.0
0.5
0.0
-0.5
0
10
20
Time (min)

Appendix 3: SASA data

SASA evaluation of the PDB deposits used in Figure 4-10.

SASA evaluation of X-ray crystal structures of native sequence RNase A

PDB ID	Resolution (Å)	Y025 (Å^2)	Y073 (Å^2)	Y076 (Å^2)	Y092 (Å^2)	Y097 (Å^2)	Y115 (Å^2)	Total (Å^2)
4J5Z	1.80	3.72	8.09	27.58	23.71	0.59	20.64	84.33
1AFU	2.00	3.07	8.96	31.54	24.54	1.55	23.73	93.39
1BEL	1.60	2.35	7.45	36.54	22.40	2.28	23.20	94.22
1BZQ	2.80	5.85	6.07	37.11	22.36	0.80	21.43	93.62
1DFJ	2.50	3.66	8.89	31.69	20.81	1.52	13.23	79.80
1EOS	2.00	3.07	8.96	33.77	25.34	1.55	22.29	94.98
1EOW	2.00	3.75	8.34	32.81	22.19	1.55	21.89	90.53
1FS3	1.40	1.48	8.09	28.50	21.47	1.31	20.64	81.49
1JN4	1.80	3.91	8.17	30.74	26.64	0.76	21.57	91.79
1JVT	2.05	2.28	7.41	31.57	23.84	1.55	20.81	87.46
1JVU	1.78	3.07	8.96	30.02	23.86	1.55	20.89	88.35
1KF5	1.15	3.00	7.37	30.81	23.86	0.80	19.58	85.42
1KF7	1.15	3.00	6.58	31.40	22.99	0.80	19.58	84.35
1O0F	1.50	5.35	8.96	30.78	23.86	0.00	20.81	89.76
1O0H	1.20	3.79	8.96	29.22	23.94	0.76	21.61	88.28
1O0M	1.50	6.14	8.96	27.71	23.88	0.76	21.61	89.06
1O0N	1.50	5.35	8.96	29.98	24.58	1.48	21.61	91.96
1O0O	1.20	2.28	8.96	29.94	26.02	0.76	21.61	89.57
1RBX	1.69	3.00	8.38	28.62	22.31	0.80	21.09	84.20
1RCA	1.90	4.30	5.99	28.54	23.75	0.80	20.30	83.68
1RHA	1.80	2.28	6.58	35.25	26.78	0.80	21.68	93.37
1RHB	1.50	3.00	5.99	34.95	23.03	0.80	23.33	91.10
1RNC	1.50	2.24	6.79	35.12	21.55	0.80	21.09	87.59
1RND	1.50	3.00	5.99	32.98	21.47	0.80	20.30	84.54
1RNM	2.00	3.54	9.61	32.40	21.47	0.59	21.61	89.22
1RNN	1.80	5.86	7.33	28.71	22.99	1.31	20.81	87.01
1RNO	1.90	3.64	7.33	29.47	23.03	0.59	22.19	86.25
1RNQ	2.00	3.54	6.61	33.05	22.48	1.31	22.91	89.90

1RNW	1.80	3.72	7.33	31.57	24.54	0.59	20.77	88.52
1RNX	1.90	2.92	7.33	30.19	20.89	0.59	24.54	86.46
1RNY	2.00	2.92	6.54	28.67	22.99	0.59	21.47	83.18
1RNZ	1.90	2.96	7.33	30.19	20.81	0.59	22.19	84.07
1ROB	1.60	3.00	9.14	32.09	20.75	0.80	21.89	87.67
1RPF	2.20	3.00	9.76	32.56	21.00	0.76	22.61	89.69
1RPG	1.40	5.06	8.38	33.64	23.78	0.80	21.09	92.75
1RPH	2.20	3.72	7.33	32.16	20.05	0.00	22.15	85.41
1RUV	1.25	2.24	9.14	28.33	21.43	0.80	20.30	82.24
1U1B	2.00	5.89	10.37	32.67	22.55	1.52	20.73	93.73
1W4O	1.60	3.15	8.17	31.50	24.66	0.76	21.61	89.85
1W4P	1.69	2.28	8.96	30.02	23.86	0.76	20.81	86.69
1W4Q	1.68	4.59	8.96	30.02	26.06	0.76	22.33	92.72
1WBU	1.90	4.59	8.96	29.98	22.36	0.76	20.09	86.74
1XPS	1.80	3.79	8.38	24.64	23.16	1.52	22.87	84.36
1XPT	1.90	4.51	7.66	34.15	25.43	0.80	20.56	93.11
1Z6D	1.54	4.59	8.96	29.98	22.34	0.76	20.89	87.52
1Z6S	1.50	3.79	9.68	31.54	23.82	0.00	21.53	90.36
2BLP	1.40	2.31	6.90	30.68	21.99	0.76	21.19	83.83
2E3W	1.05	3.15	7.62	30.68	23.56	1.55	21.19	87.75
2G8Q	1.50	3.79	8.96	31.50	23.86	0.00	21.61	89.72
2G8R	1.70	6.07	8.20	32.22	21.76	0.00	20.81	89.06
2QCA	1.33	3.75	6.69	35.55	23.27	1.52	24.28	95.06
2RNS	1.60	7.70	8.46	33.24	25.83	2.20	21.93	99.36
2W5G	1.70	2.28	8.20	29.98	26.86	0.76	22.25	90.33
2W5K	1.70	3.07	8.17	31.57	26.86	1.48	22.33	93.48
2W5L	1.70	2.28	7.41	30.78	24.51	0.76	22.33	88.07
2W5M	1.80	3.87	8.96	30.78	24.58	0.76	23.05	92.00
3D6O	1.58	5.31	8.96	31.50	23.86	0.76	21.61	92.00
3D6P	1.60	2.28	8.96	30.78	23.86	0.76	21.61	88.25
3D6Q	1.60	3.79	7.41	32.98	25.34	0.00	20.81	90.33
3D7B	1.60	3.79	8.17	32.29	24.58	0.00	22.33	91.16
3D8Y	1.72	3.79	7.41	31.46	25.34	0.00	20.85	88.85
3D8Z	1.98	4.59	8.17	30.82	23.06	0.76	20.85	88.25
3DH5	1.60	2.20	7.33	31.19	21.61	1.31	19.16	82.80

3DXG	1.39	3.03	7.41	31.57	24.58	0.00	23.81	90.40
3DXH	1.40	4.55	7.41	30.78	23.14	0.00	23.09	88.97
3MZQ	1.50	2.90	6.69	32.39	24.58	2.28	20.96	89.80
3MZR	1.50	2.31	6.69	32.52	25.38	3.07	21.68	91.65
4G8V	1.70	4.59	7.41	31.57	23.14	0.00	20.81	87.52
4G8Y	1.80	3.79	8.13	31.50	23.86	0.00	20.81	88.09
4G90	1.90	3.79	8.13	31.54	24.58	0.00	20.81	88.85
4L55	1.65	2.28	7.41	30.78	23.86	0.76	20.09	85.18
4MXF	2.25	0.00	8.17	29.26	24.68	0.76	22.36	85.23
5RSA	2.00	4.34	9.72	26.30	23.78	0.80	22.61	87.55
7RSA	1.26	3.58	9.14	30.65	23.78	0.80	21.85	89.80
9RSA	1.80	3.54	8.96	32.85	23.92	0.72	20.52	90.51
Average	1.70	3.59	8.03	31.26	23.54	0.86	21.44	88.72
Std Dev.	0.31	1.23	1.02	2.16	1.57	0.61	1.43	3.61

SASA evaluation of an NMR structure of RNase A (PDB ID: 2AAS, 32 conformers)

Conformer	Y025 (Å^2)	Y073 (Å^2)	Y076 (Å^2)	Y092 (Å^2)	Y097 (Å^2)	Y115 (Å^2)	Total (Å^2)
1	8.30	11.24	33.46	31.22	0.00	21.43	105.65
2	10.44	9.65	33.75	30.67	0.00	24.35	108.86
3	8.85	10.37	35.91	18.10	0.00	25.28	98.51
4	9.65	7.34	33.01	25.98	0.00	22.72	98.70
5	5.35	7.34	33.72	28.77	0.00	23.35	98.53
6	9.61	9.65	32.88	26.12	0.00	23.35	101.61
7	9.02	8.13	34.95	26.91	0.80	26.38	106.19
8	7.51	8.85	34.98	22.91	0.00	24.83	99.08
9	8.93	6.58	14.86	26.67	0.00	19.77	76.81
10	6.58	7.33	14.90	29.02	0.00	22.56	80.39
11	8.13	6.58	12.83	30.95	0.00	20.36	78.85
12	9.65	8.89	25.58	28.98	0.00	24.15	97.25
13	9.65	7.34	12.11	27.16	0.00	19.05	75.31
14	9.65	12.72	16.79	27.25	0.00	16.85	83.26
15	8.85	10.44	30.72	27.50	0.00	21.43	98.94
16	8.85	12.72	34.36	27.46	0.00	26.08	109.47

17	8.30	11.16	34.23	31.37	0.00	22.02	107.08
18	9.71	7.34	35.88	31.94	0.00	22.08	106.95
19	8.85	7.34	26.86	18.06	0.00	17.10	78.21
20	9.78	7.34	35.16	26.99	0.00	20.56	99.83
21	5.35	8.93	32.95	26.19	0.00	21.15	94.57
22	11.88	10.37	37.47	25.15	0.00	17.06	101.93
23	10.92	9.68	35.67	27.42	0.80	20.81	105.30
24	8.99	8.85	34.91	26.82	0.00	20.70	100.27
25	9.65	7.37	14.29	29.06	0.00	26.59	86.96
26	8.89	8.89	12.55	30.40	0.00	26.49	87.22
27	8.85	8.89	14.97	27.29	0.00	24.15	84.15
28	11.09	12.72	15.24	28.26	0.00	26.49	93.80
29	8.93	7.34	12.70	30.42	0.00	22.12	81.51
30	5.10	12.72	14.40	28.43	0.00	17.57	78.22
31	11.09	11.16	35.27	26.67	0.00	20.60	104.79
32	8.85	12.72	34.36	27.46	0.00	26.08	109.47
Average	8.91	9.31	26.93	27.43	0.05	22.30	94.93
Std. Dev.	1.59	2.01	9.70	3.16	0.20	2.94	11.19

SASA evaluation of a neutron diffraction structure of RNase A

PDB ref.	**Resolution (Å)**	**Y025 (Å²)**	**Y073 (Å²)**	**Y076 (Å²)**	**Y092 (Å²)**	**Y097 (Å²)**	**Y115 (Å²)**	**Total (Å²)**
3A1R	1.70	2.28	6.79	32.29	23.03	0.80	20.37	85.56

Appendix 4: Permissions

ELSEVIER LICENSE
TERMS AND CONDITIONS

May 05, 2018

This Agreement between University of Edinburgh -- Christopher Allan ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4338890067072
License date	Apr 30, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Molecular Structure: THEOCHEM
Licensed Content Title	On the conformation of the biotin molecule
Licensed Content Author	A.A. Strzelczyk,J.Cz. Dobrowolski,A.P. Mazurek
Licensed Content Date	May 31, 2001
Licensed Content Volume	541
Licensed Content Issue	1-3
Licensed Content Pages	8
Start Page	283
End Page	290
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	3
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figures 4, 5, 6
Title of your thesis/dissertation	Targeting tyrosine: a catch-and-release approach to protein modification
Expected completion date	Jun 2018
Estimated size (number of pages)	120
Requestor Location	University of Edinburgh Joseph Black Building David Brewster Rd EH9 3FJ Edinburgh, Midlothian EH9 3FJ United Kingdom Attn: University of Edinburgh
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD

Appendix 5: Abbreviations

Amino acid 3-letter and 1-letter codes:

Alanine	Ala	A	Arginine	Arg	R
Asparagine	Asn	N	Aspartic acid	Asp	D
Cysteine	Cys	C	Glutamic acid	Glu	E
Glutamine	Gln	Q	Glycine	Gly	G
Histidine	His	H	Isoleucine	Ile	I
Leucine	Leu	L	Lysine	Lys	K
Methionine	Met	M	Phenylalanine	Phe	F
Proline	Pro	P	Serine	Ser	S
Threonine	Thr	T	Tryptophan	Trp	W
Tyrosine	Tyr	Y	Valine	Val	V

ADC	Antibody-drug conjugate
aq.	Aqueous
Av	Avidin
BSA	Bovine serum albumin
C_H	Constant domain (heavy chain)
CID	Collision-induced dissociation
C_L	Constant domain (light chain)
COSY	Correlated Spectroscopy
CuAAC	Copper-catalysed azide-alkyne cycloaddition
CV	Column volume
Cyt. C	Cytochrome C
Da	Daltons
DCC	Dicyclohexylcarbodiimide
DCM	Dichloromethane
dd	dot density
DIC	Diisopropylcarbodiimide
DIPEA	Diisopropylethylamine
DMAP	4-Dimethylaminopyridine
DMF	Dimethylformamide
ECD	Electron-capture dissociation
EDC	1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide
EPR	Electron paramagnetic resonance
EWG	Electron-withdrawing group
FDA	Federal Drug Administration
FITC	Flourescein isothiocyanate
FT-ICR	Fourier-transform ion cyclotron resonance
GFP	Green fluorescent protein
GUI	Graphical user interface
h	Hours
HEWL	Lysozyme C

HOBt	1-Hydroxybenzotriazole
HPLC	High performance liquid chromatography
HRMS	High resolution mass spectrometry
Hz	Hertz
IUPAC	International union of pure and applied chemistry
MaxEnt	Maximum entropy deconvolution
Mb	Myoglobin
MeCN	Acetonitrile
MeOH	Methanol
min	Minutes
MS	Mass spectrometry
M_W	Molecular weight
MWCO	Molecular weight cut-off
m/z	Mass to charge ratio
NCL	Native chemical ligation
NHS	*N*-Hydroxysuccinimide
nm	Nanometres
NMR	Nuclear magnetic resonance
OC	Oxidative coupling
PABA	*para*-Aminobenzoic acid
PBD ID	Protein data bank identification code
PEG	Polyethylene glycol
ppm	Parts per million
PTAD	4-Phenyl-3*H*-1,2,4-triazoline-3,5(4*H*)-dione
PTM	Post-translational modification
RNase A	Ribonuclease A
rt	Room temperature
RuACC	Ru-catalysed azide-alkyne cycloaddition
s	Seconds
SASA	Solvent-accessible surface area
sat.	Saturated
SAv	Streptavidin
SBTI	Soybean trypsin inhibitor
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
SPAAC	Strain-promoted azide-alkyne cycloaddition
SPPS	Solid-phase peptide synthesis
SPR	Surface plasmon resonance
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
TIS	Triisopropylsilane
TLC	thin layer chromatography
TNBS	2,4,6,-Trinitrobenzenesulfonic acid
TNM	Tetranitromethane
UAA	Unnatural amino acid
UHPLC	Ultra high performance liquid chromatography
UV	Ultraviolet
VdW	Van der Waals
V_H	Variable domain (heavy chain)
V_L	Variable domain (light chain)
β- Lac	β-Lactoglobulin A

1. W. R. Algar, in *Chemoselective and Bioorthogonal Ligation Reactions: Concepts and Applications*, ed. W. R. Algar, P. E. Dawson and I. L. Medintz, Wiley-VCH Verlag GmbH & Co. KGaA, Germany, 1st edn, 2017, ch. 1, pp 3-36.
2. P. C. Fridy, Y. Y. Li, S. Keegan, M. K. Thompson, I. Nudelman, J. F. Scheid, M. Oeffinger, M. C. Nussenzweig, D. Fenyo, B. T. Chait and M. P. Rout, *Nature Methods*, 2014, **11**, 1253-1260.
3. J. N. deGruyter, L. R. Malins and P. S. Baran, *Biochemistry*, 2017, **56**, 3863-3873.
4. T. Rodrigues and G. J. L. Bernardes, *Angewandte Chemie-International Edition*, 2018, **57**, 2032-2034.
5. J. Collins, K. Kempe, P. Wilson, C. A. Blindauer, M. P. McIntosh, T. P. Davis, M. R. Whittaker and D. M. Haddleton, *Biomacromolecules*, 2016, **17**, 2755-2766.
6. P. Jonkheijm, D. Weinrich, H. Schroder, C. M. Niemeyer and H. Waldmann, *Angewandte Chemie-International Edition*, 2008, **47**, 9618-9647.
7. N. Suzuki, M. Hiraki, Y. Yamada, N. Matsugaki, N. Igarashi, R. Kato, I. Dikic, D. Drew, S. Iwata, S. Wakatsuki and M. Kawasaki, *Acta Crystallographica Section D-Biological Crystallography*, 2010, **66**, 1059-1066.
8. R. Y. Tsien, *Angewandte Chemie-International Edition*, 2009, **48**, 5612-5626.
9. P. Dedecker, F. C. De Schryver and J. Hofkens, *Journal of the American Chemical Society*, 2013, **135**, 2387-2402.
10. K. Thorn, *Molecular Biology of the Cell*, 2017, **28**, 848-857.
11. G. Zhang, S. Q. Zheng, H. P. Liu and P. R. Chen, *Chemical Society Reviews*, 2015, **44**, 3405-3417.
12. A. Gautier, A. Juillerat, C. Heinis, I. R. Correa, M. Kindermann, F. Beaufils and K. Johnsson, *Chemistry & Biology*, 2008, **15**, 128-136.
13. G. V. Los, L. P. Encell, M. G. McDougall, D. D. Hartzell, N. Karassina, C. Zimprich, M. G. Wood, R. Learish, R. F. Ohane, M. Urh, D. Simpson, J. Mendez, K. Zimmerman, P. Otto, G. Vidugiris, J. Zhu, A. Darzins, D. H. Klaubert, R. F. Bulleit and K. V. Wood, *ACS Chemical Biology*, 2008, **3**, 373-382.
14. A. Keppler, M. Kindermann, S. Gendreizig, H. Pick, H. Vogel and K. Johnsson, *Methods*, 2004, **32**, 437-444.
15. V. Liss, B. Barlag, M. Nietschke and M. Hensel, *Scientific Reports*, 2015, **5**, 17740.
16. A. Keppler, S. Gendreizig, T. Gronemeyer, H. Pick, H. Vogel and K. Johnsson, *Nature Biotechnology*, 2003, **21**, 86-89.
17. J. E. Cronan, *Journal of Biological Chemistry*, 1990, **265**, 10327-10333.
18. D. Beckett, E. Kovaleva and P. J. Schatz, *Protein Science*, 1999, **8**, 921-929.
19. J. Lotze, U. Reinhardt, O. Seitz and A. G. Beck-Sickinger, *Molecular Biosystems*, 2016, **12**, 1731-1745.
20. S. A. Walper, K. B. Turner and I. L. Medintz, in *Chemoselective and Bioorthogonal Ligation Reactions: Concepts and Applications*, ed. W. R. Algar, P. E. Dawson and I. L. Medintz, Wiley-VCH Verlag GmbH & Co. KGaA, Germany, 1st edn, 2017, ch. 7, pp 165-230.
21. S. A. Slavoff, I. Chen, Y. A. Choi and A. A. Y. Ting, *Journal of the American Chemical Society*, 2008, **130**, 1160-1162.
22. D. Schumacher, O. Lemke, J. Helma, L. Gerszonowicz, V. Waller, T. Stoschek, P. M. Durkin, N. Budisa, H. Leonhardt, B. G. Keller and C. P. R. Hackenberger, *Chemical Science*, 2017, **8**, 3471-3478.
23. D. Schumacher, J. Helma, F. A. Mann, G. Pichler, F. Natale, E. Krause, M. C. Cardoso, C. P. R. Hackenberger and H. Leonhardt, *Angewandte Chemie-International Edition*, 2015, **54**, 13787-13791.
24. S. R. Adams and R. Y. Tsien, *Nature Protocols*, 2008, **3**, 1527-1534.
25. R. C. Brewster, G. C. Gavins, B. Gunthardt, S. Farr, K. M. Webb, P. Voigt and A. N. Hulme, *Chemical Communications*, 2016, **52**, 12230-12232.
26. D. Agrawal and C. P. R. Hackenberger, *Chemistry of Organo-Hybrids: Synthesis and Characterization of Functional Nano-Objects*, ed. B. Charleux, C. Coperet and E. Lacote,John Wiley & Sons, Inc., USA, 1st edn, 2015, ch. 9, 299-348.
27. K. Lang and J. W. Chin, *Chemical Reviews*, 2014, **114**, 4764-4806.

28. J. W. Chin, *Annual Review of Biochemistry, Vol 83*, 2014, **83**, 379-408.
29. UniProtKB/TrEMBL protein database release 2017_08 statistics, http://www.ebi.ac.uk/uniprot/TrEMBLstats, (accessed 26/09, 2017).
30. A. K. Shaytan, K. V. Shaitan and A. R. Khokhlov, *Biomacromolecules*, 2009, **10**, 1224-1237.
31. D. Eisenberg, E. Schwarz, M. Komaromy and R. Wall, *Journal of Molecular Biology*, 1984, **179**, 125-142.
32. S. Miller, J. Janin, A. M. Lesk and C. Chothia, *Journal of Molecular Biology*, 1987, **196**, 641-656.
33. O. Koniev and A. Wagner, *Chemical Society Reviews*, 2015, **44**, 5495-5551.
34. P. Akkapeddi, S. A. Azizi, A. M. Freedy, P. Cal, P. M. P. Gois and G. J. L. Bernardes, *Chemical Science*, 2016, **7**, 2954-2963.
35. G. T. Hermanson, in *Bioconjugate Techniques*, ed. G. T. Hermanson, Elsevier, Amsterdam, 3rd edn, 2013, ch. 10, 395-463.
36. S. D. Fontaine, R. Reid, L. Robinson, G. W. Ashley and D. V. Santi, *Bioconjugate Chemistry*, 2015, **26**, 145-152.
37. D. G. Hoch, D. Abegg and A. Adibekian, *Chemical Communications*, 2018, **54**, 4501-4512.
38. J. R. Kim, H. W. Yoon, K. S. Kwon, S. R. Lee and S. G. Rhee, *Analytical Biochemistry*, 2000, **283**, 214-221.
39. G. T. Hermanson, in *Bioconjugate Techniques*, ed. G. T. Hermanson, Elsevier, Amsterdam, 3rd edn, 2013, ch. 3, 229-258.
40. J. M. Chalker, S. B. Gunnoo, O. Boutureira, S. C. Gerstberger, M. Fernandez-Gonzalez, G. J. L. Bernardes, L. Griffin, H. Hailu, C. J. Schofield and B. G. Davis, *Chemical Science*, 2011, **2**, 1666-1676.
41. G. J. L. Bernardes, J. M. Chalker, J. C. Errey and B. G. Davis, *Journal of the American Chemical Society*, 2008, **130**, 5052-5053.
42. C. S. Kang, N. J. Wu, Y. W. Chen, X. Sun, N. Bandara, D. J. Liu, M. R. Lewis, B. E. Rogers and H. S. Chong, *Journal of Inorganic Biochemistry*, 2016, **154**, 60-66.
43. F. X. Zeng, T. T. Qi, C. Y. Li, T. F. Li, H. L. Li, S. L. Li, L. L. Zhu and X. Y. Xu, *MedChemComm*, 2017, **8**, 1297-1302.
44. X. Chen, K. Muthoosamy, A. Pfisterer, B. Neumann and T. Weil, *Bioconjugate Chemistry*, 2012, **23**, 500-508.
45. X. Chen, L. Henschke, Q. Z. Wu, K. Muthoosamy, B. Neumann and T. Weil, *Organic & Biomolecular Chemistry*, 2013, **11**, 353-361.
46. J. M. McFarland and M. B. Francis, *Journal of the American Chemical Society*, 2005, **127**, 13490-13491.
47. N. Wang, B. Yang, C. Fu, H. Zhu, F. Zheng, T. Kobayashi, J. Liu, S. Li, C. Ma, P. G. Wang, Q. Wang and L. Wang, *Journal of the American Chemical Society*, 2018, **140**, 15, 4995-4999.
48. C. B. Rosen and M. B. Francis, *Nature Chemical Biology*, 2017, **13**, 697-705.
49. T. Arnesen, P. Van Damme, B. Polevoda, K. Helsens, R. Evjenth, N. Colaert, J. E. Varhaug, J. Vandekerckhove, J. R. Lillehaug, F. Sherman and K. Gevaert, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 8157-8162.
50. J. M. Gilmore, R. A. Scheck, A. P. Esser-Kahn, N. S. Joshi and M. B. Francis, *Angewandte Chemie-International Edition*, 2006, **45**, 5307-5311.
51. J. I. MacDonald, H. K. Munch, T. Moore and M. B. Francis, *Nature Chemical Biology*, 2015, **11**, 326-331.
52. S. Bloom, C. Liu, D. K. Kolmel, J. X. Qiao, Y. Zhang, M. A. Poss, W. R. Ewing and D. W. C. MacMillan, *Nature Chemistry*, 2018, **10**, 205-211.
53. S. S. Kulkarni, J. Sayers, B. Premdjee and R. J. Payne, *Nature Reviews Chemistry*, 2018, **2**, 0122.
54. G. Nikov, V. Bhat, J. S. Wishnok and S. R. Tannenbaum, *Analytical Biochemistry*, 2003, **320**, 214-222.
55. M. Sokolovsky, D. Harell and J. F. Riordan, *Biochemistry*, 1969, **8**, 4740-4745.
56. K. Kim, D. A. Fancy, D. Carney and T. Kodadek, *Journal of the American Chemical Society*, 1999, **121**, 11896-11897.
57. L. H. Jones, A. Narayanan and E. C. Hett, *Molecular Biosystems*, 2014, **10**, 952-969.

58. N. S. Joshi, L. R. Whitaker and M. B. Francis, *Journal of the American Chemical Society*, 2004, **126**, 15942-15943.
59. J. M. McFarland, N. S. Joshi and M. B. Francis, *Journal of the American Chemical Society*, 2008, **130**, 7639-7644.
60. H. Ban, J. Gavrilyuk and C. F. Barbas, *Journal of the American Chemical Society*, 2010, **132**, 1523-1525.
61. H. Ban, M. Nagano, J. Gavrilyuk, W. Hakamata, T. Inokuma and C. F. Barbas, *Bioconjugate Chemistry*, 2013, **24**, 520-532.
62. K. De Bruycker, S. Billiet, H. A. Houck, S. Chattopadhyay, J. M. Winne and F. E. Du Prez, *Chemical Reviews*, 2016, **116**, 3919-3974.
63. T. L. Schlick, Z. B. Ding, E. W. Kovacs and M. B. Francis, *Journal of the American Chemical Society*, 2005, **127**, 3718-3723.
64. J. Gavrilyuk, H. Ban, M. Nagano, W. Hakamata and C. F. Barbas, *Bioconjugate Chemistry*, 2012, **23**, 2321-2328.
65. M. W. Jones, G. Mantovani, C. A. Blindauer, S. M. Ryan, X. X. Wang, D. J. Brayden and D. M. Haddleton, *Journal of the American Chemical Society*, 2012, **134**, 7406-7413.
66. F. Landi, C. M. Johansson, D. J. Campopiano and A. N. Hulme, *Organic & Biomolecular Chemistry*, 2010, **8**, 56-59.
67. S. Chandrappa, K. Vinaya, T. Ramakrishnappa and K. S. Rangappa, *Synlett*, 2010, 3019-3022.
68. R. Hata, H. Nonaka, Y. Takakusagi, K. Ichikawa and S. Sando, *Chemical Communications*, 2015, **51**, 12290-12292.
69. S. Y. Jabri and L. E. Overman, *Journal of Organic Chemistry*, 2013, **78**, 8766-8788.
70. K. Rochon, A. Proteau-Gagne, P. Bourassa, J. F. Nadon, J. Cote, V. Bournival, F. Gobeil, B. Guerin, Y. L. Dory and L. Gendron, *ACS Chemical Neuroscience*, 2013, **4**, 1204-1216.
71. D. Sanfelice and P. A. Temussi, *Frontiers in Molecular Biosciences*, 2014, **1**, 1-8.
72. A. Lalatsa, V. Lee, J. P. Malkinson, M. Zloh, A. G. Schatzlein and I. F. Uchegbu, *Molecular Pharmaceutics*, 2012, **9**, 1665-1680.
73. G. Espuna, G. Arsequell, G. Valencia, J. Barluenga, M. Perez and J. M. Gonzalez, *Chemical Communications*, 2000, 1307-1308.
74. R. Subiros-Funosas, R. Prohens, R. Barbas, A. El-Faham and F. Albericio, *Chemistry-A European Journal*, 2009, **15**, 9394-9403.
75. M. Sainlos and B. Imperiali, *Nature Protocols*, 2007, **2**, 3210-3218.
76. C. L. Jaffe, H. Lis and N. Sharon, *Biochemistry*, 1980, **19**, 4423-4429.
77. S. H. L. Verhelst, M. Fonovic and M. Bogyo, *Angewandte Chemie-International Edition*, 2007, **46**, 1284-1286.
78. G. Budin, M. Moune-Dimala, G. Leriche, J. M. Saliou, J. Papillon, S. Sanglier-Cianferani, A. Van Dorsselaer, V. Lamour, L. Brino and A. Wagner, *ChemBioChem*, 2010, **11**, 2359-2361.
79. P. Dosa, I. Kronish, J. McCallum, J. Schwartz and M. C. Barden, *Journal of Organic Chemistry*, 1996, **61**, 4886-4887.
80. P. Selvam, S. K. Mohapatra, S. U. Sonavane and R. V. Jayaram, *Tetrahedron Letters*, 2004, **45**, 2003-2007.
81. S. Gowda, K. Abiraj and D. C. Gowda, *Tetrahedron Letters*, 2002, **43**, 1329-1331.
82. P. Selvam, S. U. Sonavane, S. K. Mohapatra and R. V. Jayaram, *Tetrahedron Letters*, 2004, **45**, 3071-3075.
83. H. Mutlu, C. M. Geiselhart and C. Barner-Kowollik, *Materials Horizons*, 2018, **5**, 162-183.
84. M. A. Pasha and H. M. Nanjundaswamy, *Synthetic Communications*, 2005, **35**, 897-900.
85. W. M. Koppes, J. S. Moran, J. C. Oxley and J. L. Smith, *Tetrahedron Letters*, 2008, **49**, 3234-3237.
86. A. D. Wong, T. M. Gungor and E. R. Gillies, *ACS Macro Letters*, 2014, **3**, 1191-1195.
87. X. M. Cheng, R. Y. Zhang, X. L. Cai and B. Liu, *Journal of Materials Chemistry B*, 2017, **5**, 3565-3571.
88. S. S. Liew, S. B. Du, J. Y. Ge, S. J. Pan, S. Y. Jang, J. S. Lee and S. Q. Yao, *Chemical Communications*, 2017, **53**, 13332-13335.
89. A. M. ElSohly and M. B. Francis, *Accounts of Chemical Research*, 2015, **48**, 1971-1978.
90. A. Bachi, I. Dalle-Donne and A. Scaloni, *Chemical Reviews*, 2013, **113**, 596-698.

91. H. J. C. Chen, C. M. Chang, W. P. Lin, D. L. Cheng and M. I. Leong, *ChemBioChem*, 2008, **9**, 312-323.
92. B. Balabanli, Y. Kamisaki, E. Martin and F. Murad, *Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**, 13136-13141.
93. H. Ryberg and K. Caidahl, *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*, 2007, **851**, 160-171.
94. N. Abello, B. Barroso, H. A. M. Kerstjens, D. S. Postma and R. Bischoff, *Talanta*, 2010, **80**, 1503-1512.
95. I. Y. Haddad, G. Pataki, P. Hu, C. Galliani, J. S. Beckman and S. Matalon, *Journal of Clinical Investigation*, 1994, **94**, 2407-2413.
96. A. Amoresano, G. Chiappetta, P. Pucci, M. D'Ischia and G. Marino, *Analytical Chemistry*, 2007, **79**, 2109-2117.
97. J. Guo and L. Prokai, *Journal of Mass Spectrometry*, 2012, **47**, 1601-1611.
98. M. Sokolovsky, J. F. Riordan and B. L. Vallee, *Biochemical and Biophysical Research Communications*, 1967, **27**, 20-25.
99. Y. Yang, *Journal of Chromatography A*, 2017, **1485**, 90-100.
100. M. Matos, B. Oliveira, N. Martinez-Saez, A. Guerreiro, P. M. S. D. Cal, J. Bertoldo, M. Maneiro, E. Perkins, J. Howard, M. Deery, J. M. Chalker, F. Corzana, G. Jimenez-Oses and G. J. L. Bernardes, *Journal of the American Chemical Society*, 2018, **140**, 11, 4004-4017.
101. D. G. Isom, C. A. Castaneda, B. R. Cannon and B. E. Garcia-Moreno, *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**, 5260-5265.
102. J. Y. Ng and J. W. H. Wong, *Organic & Biomolecular Chemistry*, 2015, **13**, 374-378.
103. D. L. Browne, *Angewandte Chemie-International Edition*, 2014, **53**, 1482-1484.
104. J. E. Pickett, K. Nagakura, A. R. Pasternak, S. G. Grinnell, S. Majumdar, J. S. Lewis and G. W. Pasternak, *Bioorganic & Medicinal Chemistry Letters*, 2013, **23**, 4347-4350.
105. A. Brunschweiger, P. Koch, M. Schlenk, M. Rafehi, H. Radjainia, P. Kuppers, S. Hinz, F. Pineda, M. Wiese, J. Hockemeyer, J. Heer, F. Denonne and C. E. Muller, *Bioorganic & Medicinal Chemistry*, 2016, **24**, 5462-5480.
106. T. Okazaki, K. K. Laali, S. D. Bunge and S. K. Adas, *European Journal of Organic Chemistry*, 2014, 1630-1644.
107. J. T. Fletcher and J. E. Reilly, *Tetrahedron Letters*, 2011, **52**, 5512-5515.
108. S. M. Joshi, A. de Cozar, V. Gomez-Vallejo, J. Koziorowski, J. Llop and F. P. Cossio, *Chemical Communications*, 2015, **51**, 8954-8957.
109. E. Haldon, M. C. Nicasio and P. J. Perez, *Organic & Biomolecular Chemistry*, 2015, **13**, 9528-9550.
110. J. R. Johansson, T. Beke-Somfai, A. S. Stalsmeden and N. Kann, *Chemical Reviews*, 2016, **116**, 14726-14768.
111. J. Dommerholt, F. Rutjes and F. L. van Delft, *Topics in Current Chemistry*, 2016, **374**.
112. E. Saxon and C. R. Bertozzi, *Science*, 2000, **287**, 2007-2010.
113. K. M. El Muslemany, A. A. Twite, A. M. ElSohly, A. C. Obermeyer, R. A. Mathies and M. B. Francis, *Journal of the American Chemical Society*, 2014, **136**, 12600-12606.
114. M. Berchel, J. P. Haelters, D. Afonso, A. Maroto, L. Deschamps, P. Giamarchi and P. A. Jaffres, *European Journal of Organic Chemistry*, 2014, **2014**, 1076-1083.
115. J. M. Hooker, E. W. Kovacs and M. B. Francis, *Journal of the American Chemical Society*, 2004, **126**, 3718-3719.
116. A. Anghileri, R. Lantto, K. Kruus, C. Arosio and G. Freddi, *Journal of Biotechnology*, 2007, **127**, 508-519.
117. B. M. Bizzarri, A. Martini, F. Serafini, D. Aversa, D. Piccinino, L. Botta, N. Berretta, E. Guatteo and R. Saladino, *RSC Advances*, 2017, **7**, 20502-20509.
118. J. M. Palomo, *European Journal of Organic Chemistry*, 2010, 6303-6314.
119. D. Monti, G. Ottolina, G. Carrea and S. Riva, *Chemical Reviews*, 2011, **111**, 4111-4140.
120. J. M. Hooker, PhD Thesis, University of California at Berkeley, 2007.
121. A. C. Obermeyer, J. B. Jarman, C. Netirojjanakul, K. El Muslemany and M. B. Francis, *Angewandte Chemie-International Edition*, 2014, **53**, 1057-1061.

122. R. Sangsuwan, A. C. Obermeyer, P. Tachachartvanich, K. K. Palaniappan and M. B. Francis, *Chemical Communications*, 2016, **52**, 10036-10039.
123. K. S. Palla, T. J. Hurlburt, A. M. Buyanin, G. A. Somorjai and M. B. Francis, *Journal of the American Chemical Society*, 2017, **139**, 1967-1974.
124. A. C. Obermeyer, J. B. Jarman and M. B. Francis, *Journal of the American Chemical Society*, 2014, **136**, 9572-9579.
125. P. Siejak and D. Frackowiak, *Journal of Physical Chemistry B*, 2005, **109**, 14382-14386.
126. Streptavidin Agarose 69203 Overview, http://www.merckmillipore.com/GB/en/product/Streptavidin-Agarose,EMD_BIO-69203, (accessed 21/03, 2018).
127. C. M. Dundas, D. Demonte and S. Park, *Applied Microbiology and Biotechnology*, 2013, **97**, 9343-9353.
128. R. Bielski and Z. Witczak, *Chemical Reviews*, 2013, **113**, 2205-2243.
129. J. DeChancie and K. N. Houk, *Journal of the American Chemical Society*, 2007, **129**, 5419-5429.
130. A. B. A. Jansen and P. J. Stokes, *Journal of the Chemical Society*, 1962, 4909-4914.
131. A. Gaudiano, G. Bellomonte, G. Gilardi and E. Sanzini, *Annali dell'Istituto superiore di sanita*, 1977, **13**, 773-781.
132. A. Buchynskyy, U. Kempin, S. Vogel, L. Hennig, M. Findeisen, D. Muller, S. Giesa, H. Knoll and P. Welzel, *European Journal of Organic Chemistry*, 2002, 1149-1162.
133. G. L. Waldrop, H. M. Holden and M. St Maurice, *Protein Science*, 2012, **21**, 1597-1619.
134. A. A. Strzelczyk, J. C. Dobrowolski and A. P. Mazurek, *Journal of Molecular Structure-Theochem*, 2001, **541**, 283-290.
135. L. Zhang, X. B. Hu and H. R. Li, *Journal of Physical Chemistry B*, 2008, **112**, 8779-8782.
136. S. Freitag, I. LeTrong, L. Klumb, P. S. Stayton and R. E. Stenkamp, *Protein Science*, 1997, **6**, 1157-1166.
137. A. Schmidt and M. Karas, *Journal of the American Society for Mass Spectrometry*, 2001, **12**, 1092-1098.
138. O. H. Laitinen, V. P. Hytonen, H. R. Nordlund and M. S. Kulomaa, *Cellular and Molecular Life Sciences*, 2006, **63**, 2992-3017.
139. T. Hayashi, Y. Hitomi and H. Ogoshi, *Journal of the American Chemical Society*, 1998, **120**, 4910-4915.
140. Bio-Rad tech note 1085, http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_1085.pdf, (accessed 22/09, 2017).
141. J. H. Phillips, S. A. Robrish and C. Bates, *Journal of Biological Chemistry*, 1965, **240**, 699-704.
142. P. S. Addy, S. B. Erickson, J. S. Italia and A. Chatterjee, *Journal of the American Chemical Society*, 2017, **139**, 11670-11673.
143. M. H. Kubala, O. Kovtun, K. Alexandrov and B. M. Collins, *Protein Science*, 2010, **19**, 2389-2401.
144. Schrödinger L. L. C., *The PyMOL Molecular Graphics System, Version 1.7.6.6*, 2017.
145. M. J. Betts and R. B. Russell, in *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data*, ed. M. R. Barnes, John Wiley & Sons, Inc., New Jersey, 2 edn., 2007, ch. 13, pp. 311-342.
146. D. Eisenberg, R. M. Weiss, T. C. Terwilliger and W. Wilcox, *Faraday Symposia of the Chemical Society*, 1982, 109-120.
147. J. M. Souza, E. Daikhin, M. Yudkoff, C. S. Raman and H. Ischiropoulos, *Archives of Biochemistry and Biophysics*, 1999, **371**, 169-178.
148. RCSB Protein Data Bank, https://www.rcsb.org/, (accessed 23/09, 2017).
149. B. Lee and F. M. Richards, *Journal of Molecular Biology*, 1971, **55**, 379-400.
150. E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler and J. Meiler, *Journal of Molecular Modeling*, 2009, **15**, 1093-1108.
151. A. Shrake and J. A. Rupley, *Journal of Molecular Biology*, 1973, **79**, 351-371.
152. F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander and M. Scharf, *Journal of Computational Chemistry*, 1995, **16**, 273-284.

153. J. C. Kendrew, W. Klyne, S. Lifson, T. Miyazawa, G. Nemethy, D. C. Phillips, G. N. Ramachandran and H. A. Scheraga, *The Journal of Biological Chemistry*, 1970 **245**, 6489-6497.
154. S. L. Kuan, T. Wang and T. Weil, *Chemistry-A European Journal*, 2016, **22**, 17112-17129.
155. Chemical Compuing Group, *Molecular operating Environment (MOE)*, 2012.
156. Advanced Chemistry Development Inc., *ACD/ChemSketch (Freeware)*, 2016
157. T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490-519.
158. J. Zhang, Y. W. Men, S. S. Lv, L. Yi and J. F. Chen, *Organic & Biomolecular Chemistry*, 2015, **13**, 11422-11425.
159. G. Leriche, G. Budin, L. Brino and A. Wagner, *European Journal of Organic Chemistry*, 2010, 4360-4364.
160. D. C. Laporte, K. S. Rosenthal and D. R. Storm, *Biochemistry*, 1977, **16**, 1642-1648.
161. R. Radi, *Accounts of Chemical Research*, 2013, **46**, 550-559.
162. J. A. Haas, M. A. Frederick and B. G. Fox, *Protein Expression and Purification*, 2000, **20**, 274-284.
163. N. S. Isaacs and O. H. Abed, *Tetrahedron Letters*, 1982, **23**, 2799-2802.
164. R. A. Zubarev, *Current Opinion in Biotechnology*, 2004, **15**, 12-16.
165. R. D. LeDuc, G. K. Taylor, Y. B. Kim, T. E. Januszyk, L. H. Bynum, J. V. Sola, J. S. Garavelli and N. L. Kelleher, *Nucleic Acids Research*, 2004, **32**, W340-W345.
166. L. Zamdborg, R. D. LeDuc, K. J. Glowacz, Y. B. Kim, V. Viswanathan, I. T. Spaulding, B. P. Early, E. J. Bluhm, S. Babai and N. L. Kelleher, *Nucleic Acids Research*, 2007, **35**, W701-W706.
167. M. E. Jung and J. C. Rohloff, *Journal of Organic Chemistry*, 1985, **50**, 4909-4913.
168. E. L. Jackson, *Journal of the American Chemical Society*, 1952, **74**, 837-838.
169. V. Ferro, L. Weiler and S. G. Withers, *Carbohydrate Research*, 1998, **306**, 531-538.
170. C. Goeschen, N. Wibowo, J. M. White and U. Wille, *Organic & Biomolecular Chemistry*, 2011, **9**, 3380-3385.
171. F. Bolletta, D. Fabbri, M. Lombardo, L. Prodi, C. Trombini and N. Zaccheroni, *Organometallics*, 1996, **15**, 2415-2417.
172. A. R. de la Faverie, F. Hamon, C. Di Primo, E. Largy, E. Dausse, L. Delauriere, C. Landras-Guetta, J. J. Toulme, M. P. Teulade-Fichou and J. L. Mergny, *Biochimie*, 2011, **93**, 1357-1367.
173. S. N. Mistry, J. G. Baker, P. M. Fischer, S. J. Hill, S. M. Gardiner and B. Kellam, *Journal of Medicinal Chemistry*, 2013, **56**, 3852-3865.
174. A. Eisenfuhr, P. S. Arora, G. Sengle, L. R. Takaoka, J. S. Nowick and M. Famulok, *Bioorganic & Medicinal Chemistry*, 2003, **11**, 235-249.
175. F. R. Mu, S. L. Coffing, D. J. Riese, R. L. Geahlen, P. Verdier-Pinard, E. Hamel, J. Johnson and M. Cushman, *Journal of Medicinal Chemistry*, 2001, **44**, 441-452.
176. K. Hofmann, F. M. Finn and Y. Kiso, *Journal of the American Chemical Society*, 1978, **100**, 3585-3590.